AD-A032 738 TEXAS UNIV AT AUSTIN DEPT OF ELECTRICAL ENGINEERING F/G 12/1
NONPARAMETRIC DISCRIMINATION AND DENSITY ESTIMATION.(U)
DEC 76 L P DEVROYE AF-AFOSR-2371-72

UNCLASSIFIED AFOSR-TR-76-1208 NL

1 OF 2
AD A032738

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR - TR - 76 - 1208	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): NONPARAMETRIC DISCRIMINATION AND DENSITY ESTIMATION	5. TYPE OF REPORT & PERIOD COVERED: Interim rept.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): L.P. Devroye	8. CONTRACT OR GRANT NUMBER(s): AFOSR 72-2371
9. PERFORMING ORGANIZATION NAME AND ADDRESS: University of Texas, Department of Electrical Engineering, Austin, Texas 78712	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F, 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research/NM, Bolling AFB, Washington, D.C. 20332	12. REPORT DATE: December, 1976
	13. NUMBER OF PAGES: 155
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Luc P. J. A. Devroye	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

AF-AFOSR-2371-72

165P.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

2304 A5

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Nonparametric discrimination, nonparametric density estimation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The asymptotic properties of generalized nearest neighbor rules and other nonparametric discrimination rules are investigated. The problem is to estimate an M-ary valued parameter θ if an observed random vector X and data consisting of a sequence of independent random vectors $(X_1, \theta_1), \ldots, (X_n, \theta_n)$ with the same distribution as (X, θ) are given. Conditions are given

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

401997

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract (continued)

under which the rules are asymptotically optimal.

Consistent density estimates can be used to construct asymptotically optimal rules if the distribution function of X is absolutely continuous. A detailed study is made of the pointwise, integral and uniform convergence of the Parzen-Rosenblatt and Loftsgaarden-Quesenberry density estimates.

In addition, methods of estimating the conditional probability of error with a particular data set are given. For linear discrimination rules, local rules and two-step rules, estimates are found whose performance is bounded independently of the distribution of (X, θ).

theta

NONPARAMETRIC DISCRIMINATION AND DENSITY ESTIMATION

APPROVED BY SUPERVISORY COMMITTEE:

Terry J. Wagner

L. W. Pfeifer

G. J. Lieman

J. T. Gerrard

Steven J. Marcus

NONPARAMETRIC DISCRIMINATION AND DENSITY ESTIMATION

by

LUC P J A DEVROYE, M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 1976

ACKNOWLEDGMENTS

There are no words to express my thanks to Professor T.J. Wagner. His thorough understanding of problems relating to discrimination, his critical reading of the manuscript and his innumerable suggestions have been invaluable. I am also thankful to the indefatigable Dr. C.S. Penrod for pointing out that the holdout estimate is a useful distribution-free estimate for the conditional probability of error for local rules. I also wish to thank Ms. Nancy Boster for typing part of the dissertation. But above all, this work would not have been possible were it not for the continued unselfish support of my wife Beatrice.

NONPARAMETRIC DISCRIMINATION AND DENSITY ESTIMATION

Publication No. _____

Luc P J A Devroye, Ph.D.
The University of Texas at Austin, 1976

Supervising Professor: Terry J. Wagner

The asymptotic properties of generalized nearest neighbor rules and other nonparametric discrimination rules are investigated. The problem is to estimate an M-ary valued parameter θ if an observed random vector X and data consisting of a sequence of independent random vectors $(X_1, \theta_1), \ldots, (X_n, \theta_n)$ with the same distribution as (X, θ) are given. Conditions are given under which the rules are asymptotically optimal.

Consistent density estimates can be used to construct asymptotically optimal rules if the distribution function of X is absolutely continuous. A detailed study is made of the pointwise, integral and uniform convergence of the Parzen-Rosenblatt and Loftsgaarden-Quesenberry density estimates.

In addition, methods of estimating the conditional probability of error with a particular data set are given. For linear discrimination rules, local rules and two-step rules, estimates are found whose performance is bounded independently of the distribution of (X, θ).

TABLE OF CONTENTS

Chapter 1 INTRODUCTION

The discrimination problem may be formulated as follows. The statistician collects data $(X_1, \theta_1), \ldots, (X_n, \theta_n)$, a sequence of independent identically distributed (iid) random vectors drawn from the distribution of (X, θ), a random vector independent of the data. For each $1 \le i \le n$, the observation X_i takes values in $\mathbb{R}_m$ and its state θ_i takes values in $\{1, \ldots, M\}$. The discrimination problem is that of estimating the state θ from the data and the observation X using procedures which do not require complete knowledge of the distribution of (X, θ). If $\hat{\theta}$ denotes the estimate, then a measure of the performance of the procedure, given the data $V_n = (X_1, \theta_1, \ldots, X_n, \theta_n)$, is $L_n = P\{\hat{\theta} \neq \theta \mid V_n\}$, the conditional probability of error.

If the distribution of (X, θ) is known, then the results of statistical decision theory show how to choose $\hat{\theta}$, given X, such that $P\{\hat{\theta} \neq \theta\}$ is minimal. The minimal value L^* is the Bayes probability of error and we know that $L_n \ge L^*$ for all methods of obtaining $\hat{\theta}$ from X and V_n. A discrimination rule, i.e., a sequence of methods for estimating $\hat{\theta}$ from X and V_n, is said to be asymptotically optimal if $L_n \xrightarrow{n} L^*$ in probability. The importance of asymptotically optimal discrimination rules is that, for the statistician, such rules are the only ones guaranteeing that L_n is close to L^* provided he collects enough data. One of our objectives is to show how to construct asymptotically optimal discrimination rules and to display several techniques to prove, or disprove, the asymptotic optimality of discrimination rules for certain large classes of distributions of (X, θ).

If the statistician has a fair amount of a priori knowledge about the distribution of (X, θ), then he may be able to construct a parametric

model for the distribution of (X, θ), determine the parameters in the model that best fit the data and use this particular version of the model with X to obtain an estimate $\hat{\theta}$ of θ. However, if the model is not exact, then it is usually impossible to construct an asymptotically optimal discrimination rule in this manner. In the absence of sufficient knowledge about the distribution of (X, θ) to use such a parametric model, is it still possible to construct an asymptotically optimal discrimination rule? Under some mild conditions on the distribution of (X, θ), the answer is affirmative. In Chapters 2 and 5 a rather detailed study is made of the asymptotic properties of some of the most popular rules, including some original ones. Rules that are discussed include a generalized version of the celebrated k-nearest neighbor rule, two-step rules using Parzen-Rosenblatt density estimates of Loftsgaarden-Quesenberry density estimates, and histogram type discrimination rules. Of course we are bound to repeat some results that can be found elsewhere in the literature. However, to the best of the author's knowledge, the literature lacks a technical paper or book that systematically and rigorously treats asymptotic optimality and related problems for nonparametric discrimination rules.

If the distribution of X is absolutely continuous with respect to Lebesgue measure, then it is shown in Chapter 2 how nonparametric density estimates can be employed to construct an asymptotically optimal discrimination rule. In Chapter 4 two popular density estimates, the Parzen-Rosenblatt estimate and the Loftsgaarden-Quesenberry estimate, are studied. In particular, conditions are obtained insuring the pointwise convergence and the convergence in L_r of these estimates, that are weaker than the conditions imposed by several authors over the past decade. In Chapter 3, some inequalities are developed concerning the uniform convergence of empirical measures in $\mathbb{R}^m$. These inequalities

are strong enough to prove the uniform convergence of the Parzen-Rosenblatt and Loftsgaarden-Quesenberry density estimates under the weakest conditions to date.

When the statistician has selected a discrimination rule and has collected data, he wants to know how his rule performs, that is, he would like to compute $L_n = P\{\hat{\theta} \neq \theta \mid V_n\}$. Of course, there is no way of computing L_n since the distribution of (X, θ) is unknown. The statistician is thus forced to estimate L_n from the data. It is shown in Chapter 6 that for most of the classical discrimination rules and nearly all the rules that are discussed in this paper, there exists a natural useful error estimate $\hat{L}_n$ of L_n. We display upper-bounds for $P\{|\hat{L}_n - L_n| \geq \epsilon\}$ $(\epsilon > 0)$ that do <u>not</u> depend upon the distribution of (X, θ). These bounds enable the statistician to compute how much confidence he can put in his estimate $\hat{L}_n$ even if he has no idea what the distribution of (X, θ) looks like. Error estimates that are discussed include the resubstitution estimate, the deleted estimate and the holdout estimate.

The dissertation is organized as follows. Chapters 2 and 5 treat the asymptotic optimality of nonparametric discrimination rules. To be able to read Chapter 5, a few results from Chapters 3 and 4 are needed. Chapters 3 and 4 on empirical measures and density estimation can be read separately. Chapter 6 on distribution-free error estimation, which has the highest concentration of new theorems, can be read separately too provided that the reader is willing to have a quick look at Chapter 2 for the definitions of some symbols. In Chapter 6 we have tried to provide a sound theoretical framework for further research in the relatively young area of error estimation. The proofs pertaining to the theorems of a given chapter are gathered in an appendix at the end of each chapter. The bibliography is far from exhaustive. However, for each subject, we have tried to list the leading theoretical papers,

at least one survey paper or book, and most of the technical articles in which alternative proofs or closely related theorems can be found.

Chapter 2 DISCRIMINATION

2.1 Introduction

A statistician observes a random vector X, taking values in $\mathbb{R}^m$, and wishes to estimate its state θ taking values in $\{1,\ldots,M\}$. To do so he has collected data $(X_1,\theta_1),\ldots,(X_n,\theta_n)$, a sequence of independent, identically distributed random vectors distributed as (X,θ) where

$$\begin{aligned} &\text{(a)}\quad P\{\theta = j\} = \pi_j \quad , \quad 1 \le j \le M; \text{ and} \\ &\text{(b)}\quad \text{given that } \theta = j\,, \quad X \text{ has probability measure } \mu_j \text{ on} \\ &\qquad\qquad \text{the Borel sets of } \mathbb{R}^m. \end{aligned} \tag{2.1}$$

Using (2.1) we see that the probability measure for X is

$$\mu = \sum_{j=1}^{M} \pi_j \mu_j \,.$$

We assume that (X,θ) is independent of the data.

A discrimination rule, or simply rule, is a sequence $\{\delta_n\}$ of decision functions where $\delta_n = (\delta_{n1},\ldots,\delta_{nM})$ is a Borel measurable mapping from $(\mathbb{R}^m \times \{1,\ldots,M\})^n \times \mathbb{R}^m$ to $[0,1]^M$ with the property that

$$\sum_{j=1}^{M} \delta_{nj} = 1 \,. \tag{2.2}$$

If $V_n = (X_1,\theta_1),\ldots,(X_n,\theta_n)$ denotes the data sequence and X_o is any random vector taking values in $\mathbb{R}^m$, then let θ_{V_n,X_o} be a random variable taking values in $\{1,\ldots,M\}$ whose distribution is determined by the joint distribution of V_n and X_o and by

$$P\{\theta_{V_n,X_o} = j \mid V_n, X_o\} = \delta_{nj}(V_n, X_o) \quad , \quad 1 \le j \le M \, .$$

The statistician estimates θ by $\theta_{V_n,X}$. For each n we define the <u>local conditional probability of error</u> $L_{n,X}$, the <u>conditional probability of error</u> L_n and the <u>probability of error</u> R_n by

$$L_{n,X} = P\{\theta_{V_n,X} \neq \theta \mid V_n, X\}$$

$$= 1 - E\{\delta_{n\theta}(V_n, X) \mid V_n, X\} \tag{2.3}$$

$$L_n = P\{\theta_{V_n,X} \neq \theta \mid V_n\} = E\{L_{n,X} \mid V_n\} \tag{2.4}$$

$$R_n = P\{\theta_{V_n,X} \neq \theta\} = E\{L_n\} \tag{2.5}$$

where we used the smoothing property for conditional expectations in (2.4). From the strong law of large numbers, L_n is the limiting frequency of errors when a large number of independent observations, all distributed as (X, θ), have their states estimated using δ_n and V_n.

Assume for the moment that the statistician knows the distribution (2.1). In that case he can estimate θ from X in the following fashion. Let $\delta = (\delta_1, \ldots, \delta_M)$ be a Borel measurable mapping from $\mathbb{R}^m$ to $[0,1]^M$ such that

$$\sum_{j=1}^{M} \delta_j(x) = 1 \quad , x \in \mathbb{R}^m$$

and let θ_X, his estimate of θ, be a random variable taking values in $\{1, \ldots, M\}$ whose distribution is determined by the distribution of X and

$$P\{\theta_X = j \mid X\} = \delta_j(X) \quad , \quad 1 \le j \le M \, .$$

The knowledge of (2.1) is sufficient to enable the statistician to find a δ for which $P\{\theta_X \neq \theta\}$, the probability of error with δ, is minimal among all such Borel measurable mappings from $\mathbb{R}^m$ to $[0,1]^M$. He could proceed as follows. He knows that there exist Borel measurable functions $p_1, \ldots, p_M$ from $\mathbb{R}^m$ to $[0,1]$ such that

$$p_j(X) = P\{\theta = j \mid X\} \quad , \quad 1 \leq j \leq M, \text{ ae}(\mu) \tag{2.6}$$

where ae(μ) denotes almost everywhere with respect to the measure μ. We note that $p_1, \ldots, p_M$ are unique up to a μ-null set and therefore,

$$\sum_{j=1}^{M} p_j(X) = 1 \quad \text{ae}(\mu) \ .$$

Let $\delta^* = (\delta_1^*, \ldots, \delta_M^*)$ be a Borel measurable mapping from $\mathbb{R}^m$ to $[0,1]^M$ such that

$$\sum_{j=1}^{M} \delta_j^*(x) = 1 \quad , \ x \in \mathbb{R}^m \tag{2.7}$$

and

$$\delta_j^*(x) = 0 \quad \text{whenever} \quad p_j(x) < \max_{1 \leq k \leq M} p_k(x) \ ,$$

$$1 \leq j \leq M \ , \ x \in \mathbb{R}^m \ . \tag{2.8}$$

It is clear that δ^* is not uniquely defined because δ^* depends upon the version $(p_1, \ldots, p_M)$ that is used in (2.6). Let θ_X^* be the corresponding estimate of θ, and let $L^* = P\{\theta_X^* \neq \theta\}$. It is worth noting that L^* is well-defined, that is L^*, depends upon the distribution (2.1) but not on the version $(p_1, \ldots, p_M)$ that is used in the definition of δ^*. It is not hard to see that δ^* is the best possible mapping from $\mathbb{R}^m$ to $[0,1]^M$ satisfying (2.7) because, for any other δ,

$$P\{\theta_X \neq \theta | X\} \geq P\{\theta_X^* \neq \theta | X\} \quad ae(\mu) \tag{2.9}$$

and

$$P\{\theta_X \neq \theta\} \geq P\{\theta_X^* \neq \theta\} = L^* .$$

Moreover, for any decision function δ_n, with probability one,

$$L_{n,X} \geq P\{\theta_X^* \neq \theta | X\} \quad ae(\mu) \tag{2.10}$$

and

$$L_n \geq P\{\theta_X^* \neq \theta\} = L^* .$$

The proofs of (2.9) and (2.10) are given in section 2.4. The quantity L^* is usually referred to as the Bayes probability of error and any Borel measurable mapping $\delta = (\delta_1, \ldots, \delta_M)$ from $\mathbb{R}^m$ to $[0,1]^M$ satisfying (2.7) for which the probability of error with δ is L^*, is called a Bayes decision function. In particular δ^* is a Bayes decision function.

The question that naturally arises is, does L_n converge to L^* in some probabilistic sense as n tends to infinity and how fast does it converge? We say that a rule $\{\delta_n\}$ is asymptotically optimal if

$$L_n \xrightarrow{n} L^* \text{ in probability} . \tag{2.11}$$

The next theorem deals with the connection between the convergence of δ_n to δ^* and that of L_n to L^*. Theorem 2.1 essentially implies that to construct an asymptotically optimal rule, the statistician should be looking for decision functions δ_n that approximate the unknown Bayes decision function δ^*.

Theorem 2.1. If for all $1 \leq j \leq M$ and all $x \in B$ where B is a Borel set from $\mathbb{R}^m$ with $\mu(B) = 1$,

either $\quad p_j(x) = \max_{1 \le i \le M} p_i(x)$

or $\quad \delta_{nj}(V_n, x) \xrightarrow{n} 0$ in probability (wp1)

(where $p_1, \ldots, p_M$ is a given version of (2.6)), then $L_n \xrightarrow{n} L^*$ in probability (wp1).

Let M and m be fixed throughout this dissertation. Three interesting classes of discrimination problems will be studied in more detail. If $\mu_1, \ldots, \mu_M$ are all absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^m$, we will call it a type C_1 discrimination problem. In that case, there are densities $f_1, \ldots, f_M$ corresponding to $\mu_1, \ldots, \mu_M$ such that for all Borel sets $B \subseteq \mathbb{R}^m$

$$\int_B f_i(x)dx = \mu_i(B) \quad , \quad 1 \le i \le M \ .$$

The f_i are the Radon-Nikodym derivatives of the μ_i with respect to the Lebesgue measure in $\mathbb{R}^m$. Clearly, every f_i is determined up to a μ_i-null set.

If there exists a countable set of points, say B, such that $\mu_1(B) = \ldots = \mu_M(B) = 1$, then we say that (2.1) defines a type C_2 discrimination problem. A type C_3 discrimination problem will occur if there exist μ-almost everywhere continuous versions $p_1, \ldots, p_M$ in (2.6). It turns out, as we will see in Chapter 5, that it is very easy to devise asymptotically optimal decision rules for this class of problems.

In the next two sections we will demonstrate how the problem of the development of asymptotically optimal rules $\{\delta_n\}$ for type C_1 and type C_2 discrimination problems can be reduced to the problem of estimating densities and measures.

2.2 Type C_1 Discrimination Problems

Assume that to all the μ_i in (2.1) correspond densities f_i, $1 \le i \le M$, that $M=1$ and that the statistician has a way of estimating, for all $x \in \mathbb{R}^m$, $f_1(x)$ from V_n. Let f_{1n} be a Borel measurable mapping from $\mathbb{R}^{mn} \times \mathbb{R}^m$ to $\mathbb{R}$, that is, f_{1n} is a Borel measurable function of $X_1, \ldots, X_n$ and x. We say that the sequence of estimates $\{f_{1n}\}$ of f_1 is weakly (strongly) consistent on B (where B is a Borel set from $\mathbb{R}^m$) if

$$f_{1n}(x) \xrightarrow{n} f_1(x) \quad \text{in probability (wp1), all } x \in B. \tag{2.12}$$

If $M>1$, it is only natural to estimate the f_j using only those (X_i, θ_i) from V_n for which $\theta_i = j$. Therefore, define the random variables $N_{1n}, \ldots, N_{Mn}$ by

$$N_{jn} = \sum_{i=1}^{n} I_{\{\theta_i = j\}} \quad , 1 \le j \le M \, , \tag{2.13}$$

so that N_{jn} is the number of observations in V_n that have state j. It is clear that

$$\sum_{j=1}^{M} N_{jn} = n \, , \text{all } n. \tag{2.14}$$

If we are going to use the $f_{jN_{jn}}$ as estimates of f_j, $1 \le j \le M$, we can ask ourselves if the sequences $\{f_{jN_{jn}}\}$ inherit the nice consistency properties given in (2.12). This question will be treated further on.

If $f = \sum_{j=1}^{M} \pi_j f_j$, then it is not hard to see that a version $(p_1, \ldots, p_M)$ of (2.6) is given by

$$p_j(x) = \begin{cases} \pi_j f_j(x)/f(x) & \text{, if } f(x) > 0 \\ 0 & \text{, if } f(x) = 0 \end{cases} \quad , 1 \le j \le M. \tag{2.15}$$

From (2.7),(2.8),we know that if $\delta^*=(\delta^*_1,\ldots,\delta^*_M)$ is a Borel measurable mapping from $\mathbb{R}^m$ to $[0,1]^M$ satisfying (2.7) and

$$\delta^*_j(x) = 0 \text{ if } \pi_j f_j(x) < \max_{1\le i\le M} \pi_i f_i(x) \ , 1\le j\le M, \tag{2.16}$$

then δ^* is a Bayes decision function for the discrimination problem defined by (2.1).To obtain a decision function that is close to δ^*,we could thus try to estimate the $\pi_j f_j$ and use these estimates in (2.16). We will of course use the density estimates $f_{jN_{jn}}$,$1\le j\le M$, to approximate the f_j.We estimate the π_j by π_{jn},$1\le j\le M$,where

$$\pi_{jn} = N_{jn}/n \ , 1\le j\le M. \tag{2.17}$$

Let δ_n be a decision function satisfying

$$\delta_{nj}(V_n,x) = 0 \text{ if } \pi_{jn} f_{jN_{jn}}(x) < \max_{1\le i\le M} \pi_{in} f_{iN_{in}}(x) \ ,$$
$$1\le j\le M, x\in\mathbb{R}^m. \tag{2.18}$$

We show that the following is true.

Theorem 2.2. Let $\{f_{jn}\}$ be a weakly (strongly) consistent sequence of estimates of f_j on B for all $1\le j\le M$,and let $\mu(B)=1$.If $\{\delta_n\}$ is a rule satisfying (2.18),then $L_n \xrightarrow{n} L^*$ in probability (wp1).

Theorem 2.2 states that if we have a sequence of estimates of f_j $(1\le j\le M)$ that is weakly consistent ae(μ),then we can construct an asymptotically optimal decision rule $\{\delta_n\}$.One way of constructing such a rule is as in (2.17),(2.18).We will refer to decision rules of this type as two-step rules.

Several types of weakly and strongly consistent sequences of density estimates are studied in chapter 4.In chapter 5,the corresponding two-step rules are discussed and some other decision rules

are developed that are asymptotically optimal for all type C_1' discrimination problems, that is, the class of problems for which there exist μ-almost everywhere continuous densities $f_i, 1 \le i \le M$. We note here that type C_1' problems are necessarily type C_1 and type C_3 problems but that every problem that is type C_1 and type C_3 is not in general a type C_1' problem.

2.3 Type C_2 Discrimination Problems

Let there exist a set

$$B = \bigcup_{k=1}^{\infty} \{x_k\}$$

with $\mu(B)=1$. Every μ_j is with probability one characterizable by the probability distribution $(m_{j1}, m_{j2}, \ldots)$ where

$$m_{jk} = P\{X=x_k \mid \theta=j\} \ , 1 \le j \le M, k=1,2,\ldots \ . \tag{2.19}$$

Let $m_k = P\{X=x_k\}, k=1,2,\ldots$ and note that

$$P\{X=x_k\} = \sum_{j=1}^{M} P\{\theta=j\} P\{X=x_k \mid \theta=j\} = \sum_{j=1}^{M} \pi_j m_{jk} \ , \quad k=1,2,\ldots \ . \tag{2.20}$$

Further, any Borel measurable function $p_j: \mathbb{R}^m \to [0,1]$ coinciding on B with (2.21) is a version of (2.6) :

$$p_j(x_k) = P\{\theta=j \mid X=x_k\} = \begin{cases} \pi_j m_{jk}/m_k & \text{if } m_k > 0 \\ 0 & \text{if } m_k = 0 \end{cases} \ , 1 \le j \le M, k=1,2,\ldots \ . \tag{2.21}$$

From section 2.1 we know that any Borel measurable function $\delta^* = (\delta_1^*, \ldots, \delta_M^*)$ from $\mathbb{R}^m$ to $[0,1]$ satisfying (2.7) and for which

$$\delta_j^*(x_k) = 0 \text{ if } \pi_j m_{jk} < \max_{1 \le i \le M} \pi_i m_{ik} \ , 1 \le j \le M, k=1,2,\ldots \tag{2.22}$$

is a Bayes decision function for the discrimination problem defined by (2.1). Our aim is to replace the $\pi_j m_{jk}$ in (2.22) by suitable estimates so that the newly obtained decision rule is asymptotically optimal.

Let $N_{1n},\ldots,N_{Mn}$ be the number of occurrences of the states $1,\ldots,M$ in the data $(X_1,\theta_1),\ldots,(X_n,\theta_n)$ (see (2.13)). Define

$$N_{jn}^k = \sum_{i=1}^{n} I_{\{\theta_i=j\}} I_{\{X_i=x_k\}}, 1\le j\le M, k=1,2,\ldots . \tag{2.23}$$

N_{jn}^k is thus the number of (X_i,θ_i) with $X_i=x_k$ and $\theta_i=j$. Obviously,

$$\sum_{k=1}^{\infty} N_{jn}^k = N_{jn}, 1\le j\le M .$$

The natural estimate of $\pi_j m_{jk}$ is N_{jn}^k/n. Therefore, let δ_n be a decision function satisfying

$$\delta_{nj}(V_n,x_k) = 0 \text{ if } N_{jn}^k < \max_{1\le i\le M} N_{in}^k, 1\le j\le M, k=1,2,\ldots . \tag{2.24}$$

Notice that (2.24) does not put any restrictions on δ_n outside B. The following theorem can be proved.

Theorem 2.3. For any decision rule satisfying (2.24) for all n, $L_n \overset{n}{\to} L^*$ wpl. In fact, for every $\epsilon>0$, there exist constants $K_1>0$ and $K_2>0$ depending upon $\pi_1,\ldots,\pi_M,\mu_1,\ldots,\mu_M$ and ϵ such that

$$P\{L_n-L^*>\epsilon\} \le K_1 e^{-K_2 n} . \tag{2.25}$$

The first part of theorem 2.3 can be proved by means of theorem 2.1 (which in turn uses lemma 2.3). We have added an alternative proof which has the advantage of being more direct and of providing more information concerning the rate of convergence of L_n to L^*. Using the Borel-Cantelli lemma, the bound (2.25) is strong enough to prove that $L_n \overset{n}{\to} L^*$ wpl.

If the $\mu_1, \ldots, \mu_M$ are not atomic measures, then one can partition $\mathbb{R}^m$ into a countable number of sets, consider each set as one point in a new space and use a decision rule satisfying (2.24) for all n in this new space. The discrimination rule so obtained is called a histogram type discrimination rule.

2.4 Proofs

Proof of (2.9) and (2.10)

Given δ_n, note that ae(μ),

$$L_{n,X} = 1 - E\{\delta_{n\theta}(V_n,X) \mid V_n,X\}$$

$$= \sum_{j=1}^{M} p_j(X) - \sum_{j=1}^{M} E\{\delta_{nj}(V_n,X) I_{\{\theta=j\}} \mid V_n,X\}$$

$$= \sum_{j=1}^{M} \left(p_j(X) - \delta_{nj}(V_n,X)\, P\{\theta=j \mid V_n,X\} \right)$$

$$= \sum_{j=1}^{M} \left(p_j(X) - \delta_{nj}(V_n,X)\, P\{\theta=j \mid X\} \right)$$

$$= \sum_{j=1}^{M} p_j(X) \left(1 - \delta_{nj}(V_n,X) \right). \tag{2.26}$$

Similarly, for any Borel measurable mapping δ from $\mathbb{R}^m$ to $[0,1]^M$ with

$$\sum_{j=1}^{M} \delta_j(x) = 1 \quad , \quad x \in \mathbb{R}^m \, , \tag{2.27}$$

$$P\{\theta_X \neq \theta \mid X\} = \sum_{j=1}^{M} p_j(X) \left(1 - \delta_j(X) \right) \quad \text{ae}(\mu).$$

Let us therefore define

$$L_n(x) = \sum_{j=1}^{M} p_j(x)\Big(1 - \delta_{nj}(V_n, x)\Big) \tag{2.28}$$

$$L\ (x) = \sum_{j=1}^{M} p_j(x)\Big(1 - \delta_j(V_n, x)\Big) \tag{2.29}$$

and

$$L^*(x) = \sum_{j=1}^{M} p_j(x)\Big(1 - \delta_j^*(V_n, x)\Big) \tag{2.30}$$

where δ^* is the Bayes decision function defined by (2.7) and (2.8). We see that $L_n(X) = L_{n,X}$ ae(μ), $P\{\theta_X \neq \theta \mid X\} = L(X)$ ae(μ) and $P\{\theta_X^* \neq \theta \mid X\} = L^*(X)$ ae(μ) .

To show (2.10), note that ae(μ), wp1,

$$L_{n,X} = L_n(X) = \sum_{j=1}^{M} p_j(X)\Big(1 - \delta_{nj}(V_n, X)\Big)$$

$$\geq \sum_{j=1}^{M} p_j(X) - \max_{1 \leq k \leq M} p_k(X)$$

$$= L^*(X) = P\{\theta_X^* \neq \theta \mid X\} .$$

Taking conditional expectations yields, wp1,

$$L_n = E\{L_{n,X} \mid V_n\} \geq P\{\theta_X^* \neq \theta\} = L^* .$$

In a similar fashion we can prove (2.9).

Q.E.D.

Proof of Theorem 2.1

Let $(p_1, \ldots, p_M)$ be a given version of (2.6), and let

$$J(x) = \{i: 1 \le i \le M \ ; \ p_i(x) = \max_{1 \le j \le M} p_j(x)\} . \tag{2.31}$$

Then from (2.28-2.31), we have

Lemma 2.1.

$$L_n(x) - L^*(x) = \sum_{j=1}^{M} \left(\max_{1 \le i \le M} p_i(x) - p_j(x) \right) \delta_{nj}(V_n, x)$$

Proof:

From (2.28) and (2.30) we have

$$L_n(x) - L^*(x) = \sum_{j=1}^{M} p_j(x) \left(\delta_j^*(x) - \delta_{nj}(V_n, x) \right)$$

$$= \max_{1 \le i \le M} p_i(x) \left(1 - \sum_{j \in J(x)} \delta_{nj}(V_n, x) \right)$$

$$- \sum_{j \notin J(x)} p_j(x) \ \delta_{nj}(V_n, x)$$

$$= \sum_{j \notin J(x)} \left(\max_{1 \le i \le M} p_i(x) - p_j(x) \right) \delta_{nj}(V_n, x) .$$

Lemma 2.1 now follows from the definition of $J(x)$.

Q.E.D.

Lemma 2.2.

If $\delta_{nj}(V_n, x) \xrightarrow{n} 0$ in probability (wp1) ae(μ) for all $j \notin J(x)$, then $L_n(x) - L^*(x) \xrightarrow{n} 0$ in probability (wp1) ae(μ).

Proof:

Lemma 2.2 follows from lemma 2.1.

Q.E.D.

We will make extensive use of the following lemma.

Lemma 2.3.

Let g_n be a Borel measurable function of V_n and x taking values in $[0,1]$ where $x \in \mathbb{R}^m$ and $n = 1,2,\ldots$. If

$$g_n(V_n,x) \overset{n}{\to} 0 \text{ in probability (wp1) ae}(\mu)$$

then

$$\int_{\mathbb{R}^m} g_n(V_n,x)\mu(dx) = E\{g_n(V_n,X) \mid V_n\}$$

$$\overset{n}{\to} 0 \text{ in probability (wp1).}$$

Proof:

For the in probability part, we argue as follows. Let $\epsilon > 0$ be arbitrary. Then

$$P\left\{\int_{\mathbb{R}^m} g_n(V_n,x)\mu(dx) \geq \epsilon\right\}$$

$$\leq \frac{1}{\epsilon} E\left\{\int_{\mathbb{R}^m} g_n(V_n,x)\mu(dx)\right\}$$

$$= \frac{1}{\epsilon}\int_{\mathbb{R}^m} E\{g_n(V_n,x)\}\mu(dx)$$

$$\overset{n}{\to} \frac{1}{\epsilon}\int_{\{x:\, g_n(V_n,x) \overset{n}{\not\to} 0 \text{ in probability}\}} \mu(dx) = 0$$

by Markov's inequality, Tonelli's theorem and two applications of the Lebesgue dominated convergence theorem. For the wp1 part of the theorem, we can argue as in the proof of theorem A of Glick (1974).

Q.E.D.

As a corollary of lemma 2.3, we have the following lemma.

Lemma 2.4.

If $L_n(x) - L^*(x) \xrightarrow{n} 0$ in probability (wp1) ae(μ), then $L_n - L^* \xrightarrow{n} 0$ in probability (wp1).

Proof:

Notice that

$$L_n - L^* = E\{L_n(X) - L^*(X) \mid V_n\}$$

and that by (2.28), (2.30), L_n and L^* are Borel measurable functions of V_n and x taking values in [0,1] ae(μ), where the set

$$\{x: L_n(x) \notin [0,1] \text{ or } L^*(x) \notin [0,1]\}$$

$$\subset \{x: \sum_{j=1}^{M} p_j(x) \neq 1\}$$

which is a μ-null set not depending upon V_n. It is easy to see that lemma 2.3 remains valid for this case.

Q.E.D.

Lemmas 2.2 and 2.4 together imply theorem 2.1.

Proof of Theorem 2.2

The following three lemmas are needed to prove theorem 2.2.

Lemma 2.5.

Let $\{f_{jn}\}$ be a weakly (strongly) consistent sequence of estimates of f_j at x, $1 \leq j \leq M$. Let $\pi_{1n}, \ldots, \pi_{Mn}$ be defined by (2.17). Then, for all $1 \leq j \leq M$,

$$|\pi_{jn} f_{jN_{jn}}(x) - \pi_j f_j(x)| \xrightarrow{n} 0 \text{ in probability (wp1)}.$$

Proof:

Let $\epsilon > 0$ be arbitrary and let $j = 1$ without loss of generality. If $\pi_1 = 0$, then $\pi_{1n} = 0$ wpl for all n and lemma 2.5 follows. So we do assume that $\pi_1 > 0$.

Recall that for any two sequences of random variables $Y_1,\ldots,Y_n,\ldots$ and $Z_1,\ldots,Z_n,\ldots$, if $Y_n \overset{n}{\to} Y$ in probability (wpl) and $Z_n \overset{n}{\to} Z$ in probability (wpl), where Y and Z are arbitrary random variables. Therefore, we only need to show that $|f_{1n}(x) - f_1(x)| \overset{n}{\to} 0$ in probability (wpl) implies that $|f_{1N_{1n}}(x) - f_1(x)| \overset{n}{\to} 0$ in probability (wpl), because we know by the strong law of large numbers that $|\pi_{1n} - \pi_1| \overset{n}{\to} 0$ wpl.

Then it is obvious that

$$P\{|f_{1N_{1n}}(x) - f_1(x)| \geq \epsilon\}$$

$$\leq P\{N_{1n} < \frac{n\pi_1}{2}\} + P\{N_{1n} \geq \frac{n\pi_1}{2};\ |f_{1N_{1n}}(x) - f_1(x)| \geq \epsilon\}$$

$$\leq P\left\{\frac{N_{1n} - E\{N_{1n}\}}{n} < -\frac{\pi_1}{2}\right\} + \inf_{k \geq \frac{n\pi_1}{2}} P\{|f_{1k}(x) - f_1(x)| \geq \epsilon\}$$

$$\overset{n}{\to} 0$$

by the weak law of large numbers and by our assumption. For the wpl part of the theorem, note that

$$P\{\bigcup_{k \geq n} \{|f_{1N_{1k}}(x) - f_1(x)| \geq \epsilon\}\}$$

$$\leq P\{N_{1n} < \frac{n\pi_1}{2}\} + P\{N_1 \geq \frac{n\pi_1}{2};\ \bigcup_{k \geq n} \{|f_{1N_{1k}}(x) - f_1(x)| \geq \epsilon\}$$

$$\leq P\{|\pi_{1n} - \pi_1| \geq \frac{\pi_1}{2}\} + P\{\bigcup_{k \geq n\pi_1/2} \{|f_{1k}(x) - f_1(x)| \geq \epsilon\}\}$$

$\xrightarrow{n} 0$

by the weak law of large numbers for π_{1n} and by our assumption.

With the aid of lemma 2.5, the trivial lemma 2.6 given below, (2.15) and lemma 2.2, we are in a position to deduce lemma 2.7.

Lemma 2.6.

If $\{\delta_n\}$ is a decision rule satisfying (2.18) and if for all $1 \le j \le M$,

$$|\pi_{jn} f_{jN_{jn}}(x) - \pi_j f_j(x)| \xrightarrow{n} 0 \text{ in probability (wpl)},$$

then $\delta_{nj}(V_n, x) \xrightarrow{n} 0$ in probability (wpl) for all j with $\pi_j f_j(x) < \underset{1 \le i \le M}{\text{Max}} \pi_i f_i(x)$.

Proof:

We prove the convergence in probability version of lemma 2.6. The wpl version is proved similarly. Let $x \in \mathbb{R}^m$ and let $J(x) = \{j_1, \ldots, j_N\} \subseteq \{1, \ldots, M\}$ be as in (2.31) where $(p_1, \ldots, p_M)$ is defined by (2.15). Let

$$d = \inf_{\substack{j \in J(x) \\ k \notin J(x)}} \left(\pi_j f_j(x) - \pi_k f_k(x)\right).$$

By definition of $J(x)$, we know that $d > 0$. Let $\epsilon > 0$ be arbitrary and let $j \notin J(x)$. Then

$$P\{\delta_{nj}(V_n, x) > \epsilon\} \le P\{\pi_{jn} f_{jN_{jn}}(x) = \underset{1 \le i \le M}{\text{Max}} \pi_{in} f_{iN_{in}}(x)\}$$

$$\le P\left\{\bigcup_{k=1}^{M} \{|\pi_{kn} f_{kN_{kn}}(x) - \pi_k f_k(x)| \ge d/2\}\right\} \xrightarrow{n} 0$$

since $|\pi_{jn} f_{jN_{jn}}(x) - \pi_j f_j(x)| \xrightarrow{n} 0$ in probability for all $1 \le j \le M$.

Q.E.D.

Lemma 2.7.

Let $\{f_{jn}\}$ be a weakly (strongly) consistent sequence of estimates of f_j on B for all $1 \le j \le M$ and let $\mu(B) = 1$. If $\{\delta_n\}$ is defined by (2.18) for all n, then

$$L_n(x) - L^*(x) \overset{n}{\to} 0 \text{ in probability (wp1), all } x \in B$$

where L_n and L^* are defined by (2.28), (2.30) and depend upon the version $(p_1,\ldots,p_M)$ of (2.6) that is defined by (2.15) and the given densities $f_1,\ldots,f_M$.

Theorem 2.2 follows from lemmas 2.7 and 2.4.

Proof of Theorem 2.3

By theorem 2.1 it suffices to show for all $x \in B$ and all $j \in \{1,\ldots,M\}$ with

$$p_j(x) < \underset{1\le i\le M}{\text{Max}}\ p_i(x)$$

that $\delta_{nj}(V_n,x) \overset{n}{\to} 0$ wp1.

Let $x_k \in B$ and let $j \notin J(x_k)$. We know that

$$\pi_j m_{jk} = \underset{1\le i\le M}{\text{Max}}\ \pi_i m_{ik} - d$$

for some $d > 0$. Next, let $\epsilon > 0$. It is clear that

$$P\{\delta_{nj}(V_n,x_k) \ge \epsilon\}$$

$$\le P\left\{\bigcup_{i=1}^{M}\left\{\left|\frac{N_{in}^k}{n} - \pi_i m_{ik}\right| \ge \frac{d}{2}\right\}\right\}$$

$$\le 2Me^{-2n(d/2)^2}$$

by Hoeffding's inequality (Hoeffding, 1963). Clearly,

$\sum_{n=1}^{\infty} P\{ \delta_{nj}(V_n, x_k) \geq \epsilon \} < \infty$ so that $\delta_{nj}(V_n, x_k) \overset{n}{\to} 0$ wp1 by the Borel-Cantelli lemma.

Q.E.D.

Alternative Proof of Theorem 2.3

Let $\epsilon > 0$ be arbitrary. Define for any $\{i_1, \ldots, i_N\} \subseteq \{1, \ldots, M\}$ with $1 \leq i_1 < \ldots < i_N \leq M$, $1 \leq N \leq M$, $b > 0$ and $a > 0$

$$B_{i_1,\ldots,i_N} = \{x_k : p_{i_1}(x_k) = \ldots = p_{i_N}(x_k) > \sup_{j \notin \{i_1,\ldots,i_N\}} p_j(x_k) \; ; \; x_k \in B\}$$

$$C^b_{i_1,\ldots,i_N} = \{x_k : p_{i_1}(x_k) = \ldots = p_{i_N}(x_k) > \sup_{j \notin \{i_1,\ldots,i_N\}} p_j(x_k) + b \; ; \; x_k \in B\}$$

$$D^a = \{x_k : m_k \geq a \; ; \; x_k \in B\}$$

and

$$F = \bigcup_{N=1}^{M} \bigcup_{\{i_1,\ldots,i_N\} \subseteq \{1,\ldots,M\}} C^b_{i_1,\ldots,i_N} \cap D^a .$$

Because

$$B = \bigcup_{N=1}^{M} \bigcup_{\{i_1,\ldots,i_N\} \subseteq \{1,\ldots,M\}} B_{i_1,\ldots,i_N}$$

we note that

$$B \cap F^c = \bigcup_{N=1}^{M} \bigcup_{\{i_1,\ldots,i_N\} \subseteq \{1,\ldots,M\}} B_{i_1,\ldots,i_N} \cap (C^b_{i_1,\ldots,i_N})^c \cup D^a$$

where $(\cdot)^c$ denotes the complement of a set. It is possible to choose

a and b small enough so that

$$P\{X \in F^c\} < \epsilon/2 .$$

Therefore, we have, ae(μ) and wpl,

$$L_n(X) - L^*(X) \le I_{\{X \in F^c\}} + I_{\{X \in F\}}\left(L_n(X) - L^*(X)\right)$$

and thus, wpl,

$$L_n - L^* \le P\{X \in F^c\} + E\{I_{\{X \in F\}}\left(L_n(X) - L^*(X)\right) \mid V_n\}$$

and

$$P\{L_n - L^* \ge \epsilon\} \le P\left\{E\{I_{\{X \in F\}}\left(L_n(X) - L^*(X)\right) \mid V_n\} > \epsilon/2\right\}$$

$$\le \frac{2}{\epsilon} E\{I_{\{X \in F\}}\left(L_n(X) - L^*(X)\right)\}$$

$$\le \frac{2}{\epsilon} \sum_{N=1}^{M} \sum_{\{i_1,\ldots,i_N\} \subseteq \{1,\ldots,M\}} \sup_{x \in D^a \cap C^b_{i_1,\ldots,i_N}} E\{L_n(x) - L^*(x)\}$$

$$\le \frac{2}{\epsilon} \sum_{N=1}^{M} \sum_{\{i_1,\ldots,i_N\} \subseteq \{1,\ldots,M\}} \sup_{x \in D^a \cap C^b_{i_1,\ldots,i_N}} \sum_{j \notin \{i_1,\ldots,i_N\}} E\left\{\left|(p_{i_1}(x) - p_j(x))\, \delta_{nj}(V_n, x)\right|\right\}$$

$$\le \frac{2}{\epsilon} \sum_{N=1,}^{M} \sum_{\{i_1,\ldots,i_N\} \subseteq \{1,\ldots,M\}} N \sup_{x \in D^a \cap C^b_{i_1,\ldots,i_N}} \sum_{j \notin \{i_1,\ldots,i_N\}} P\{\delta_{nj}(V_n, x) > 0\}$$

$$\le \frac{2}{\epsilon} \sum_{N=1,}^{M} \sum_{\{i_1,\ldots,i_N\} \subseteq \{1,\ldots,M\}} 2NM\, e^{-2n(ab/2)^2}$$

$$\le \frac{2}{\epsilon} \sum_{N=1,}^{M} \binom{M}{N} 2NM\, e^{-na^2b^2/2}$$

$$\leq \frac{1}{\epsilon}\ 2^M M^2 e^{-na^2b^2/2}$$

where we used Chebyshev's inequality, lemma 2.1, the inequality used in the first proof of theorem 2.3 and the fact that if x_k belongs to $C^b_{i_1,\ldots,i_N} \cap D^a$, then $\pi_{i_1} m_{i_1 k} \geq \pi_j m_{jk} + ab$ for all $j \notin \{i_1, \ldots, i_N\}$. This proves (2.25). Theorem 2.3 follows by the Borel-Cantelli lemma and the arbitrariness of ϵ.

Q.E.D.

Chapter 3 EMPIRICAL MEASURES

3.1 Introduction

Let $\mathcal{B}$ be the Borel sets of $\mathbf{R}^m$ and let $X_1,\ldots,X_n$ be independent identically distributed random vectors with values in $\mathbf{R}^m$ and with a common probability measure μ. The empirical measure μ_n for $X_1,\ldots,X_n$ is defined by

$$\mu_n(A) = \Big(\sum_{i=1}^{n} I_{\{X_i \in A\}}\Big)/n \quad , A \in \mathcal{B} . \tag{3.1}$$

In this section we study the closeness of $\mu_n(A)$ to $\mu(A)$ and, in particular, we obtain explicit upper bounds for

$$\sup_{A \in \mathcal{C}} P\{\,|\mu_n(A)-\mu(A)| \geq \epsilon\,\} \tag{3.2}$$

and

$$P\{\sup_{A \in \mathcal{C}} |\mu_n(A)-\mu(A)| \geq \epsilon\,\} \tag{3.3}$$

where $\epsilon > 0$ and where $\mathcal{C}$ is a given subclass of Borel sets.

By Hoeffding's inequality (Hoeffding, 1963) we already know that

$$\sup_{A \in \mathcal{B}} P\{\,|\mu_n(A)-\mu(A)| \geq \epsilon\,\} \leq 2e^{-2n\epsilon^2} . \tag{3.4}$$

Further, if $\mathcal{C} \subseteq \mathcal{B}$ is such that

$$\sup_{A \in \mathcal{C}} \mu(A) = g \leq 1/2$$

then, by Bennett's inequality (Bennett, 1962), we conclude that for all $A \in \mathcal{C}$

$$P\{\,|\mu_n(A)-\mu(A)| \geq \epsilon\,\} \leq 2e^{-n\epsilon((1+g/\epsilon)\ln(1+\epsilon/g)-1)} .$$

From $\ln(1+a/b) \geq 2a/(2b+a)$ for all $a>0, b>0$,

$$\sup_{A\in\mathcal{C}} P\{|\mu_n(A)-\mu(A)|\geq\epsilon\}\leq 2e^{-n\epsilon^2/(2g+\epsilon)}. \qquad (3.5)$$

Let us turn now to the study of upper bounds for (3.3), postponing for the time being the question of the measurability of $\sup_{A\in\mathcal{C}}|\mu_n(A)-\mu(A)|$. We remark that for absolutely continuous measures μ,

$$\sup_{A\in\mathcal{B}}|\mu_n(A)-\mu(A)| = 1 \text{ wpl} \qquad (3.6)$$

so that it makes sense to restrict ourselves to proper subsets $\mathcal{C}$. The best known result is perhaps the Glivenko-Cantelli lemma (see Loeve, 1963) which states that

$$\sup_{A\in\mathcal{C}}|\mu_n(A)-\mu(A)| \overset{n}{\to} 0 \text{ wpl} \qquad (3.7)$$

where $m=1$ and $\mathcal{C}=\{(-\infty,x]\,|\,x\in\mathbf{R}\}$. This result was later generalized to $\mathbf{R}^m$ by Wolfowitz (1960). If $\mathcal{K}^\ell$ is the class of all sets from $\mathbf{R}^m$ that are obtained by intersecting ℓ closed linear halfspaces (i.e. sets of the form $a_1x_1+\ldots+a_mx_m\geq a_{m+1}$ where $a_i,x_i\in\mathbf{R}$, $1\leq i\leq m$ and $a_{m+1}\in\mathbf{R}$) then Rao (1962) shows that

$$\sup_{A\in\mathcal{K}^\ell}|\mu_n(A)-\mu(A)| \overset{n}{\to} 0 \text{ wpl}. \qquad (3.8)$$

On the other hand, there are μ for which

$$\sup_{A\in\bigcup_{\ell=1}^{\infty}\mathcal{K}^\ell}|\mu_n(A)-\mu(A)| = 1 \text{ wpl}. \qquad (3.9)$$

Another interesting class of Borel sets from $\mathbf{R}^m$ is the class $\mathcal{C}_C$ of Borel measurable convex sets from $\mathbf{R}^m$. Rao (1962) shows that there are μ for which

$$\sup_{A\in\mathcal{C}_C}|\mu_n(A)-\mu(A)| = 1 \text{ wpl}$$

but that if every A in $\mathcal{C}_C$ has a μ-null boundary (which is always the case if μ is absolutely continuous with respect to Lebesgue measure), then

$$\sup_{A\in\mathcal{C}_C} |\mu_n(A)-\mu(A)| \overset{n}{\to} 0 \text{ wpl}$$

if and only if $\mu_n(A) \overset{n}{\to} \mu(A)$ wpl for every A with μ(boundary A)=0.

Our main objective is not so much to obtain new results of the type (3.7),(3.8) but to find explicit upper bounds for (3.3) for some interesting subclasses $\mathcal{C}$. However, we will display upper bounds that are strong enough to imply (3.7) and (3.8) by the Borel-Cantelli lemma. Most inequalities encountered in the literature regarding (3.3) deal with the class of sets $(-\infty,x], x\in\mathbb{R}^m$. The interest in this class $\mathcal{C}_0$ is that if F and F_n are the distribution functions corresponding to μ and μ_n, then

$$D_n \triangleq \sup_{x\in\mathbb{R}^m} |F_n(x)-F(x)| = \sup_{A\in\mathcal{C}_0} |\mu_n(A)-\mu(A)| . \quad (3.10)$$

For the case m=1, Dvoretzky, Kiefer and Wolfowitz (1956) showed that there exists a universal constant C>0 such that for all $\epsilon>0$

$$P\{D_n \geq \epsilon\} \leq C e^{-2n\epsilon^2}. \quad (3.11)$$

For m>1, Kiefer and Wolfowitz (1958) later showed that there exist constants $C_1>0, C_2>0$, both depending upon m, such that

$$P\{D_n \geq \epsilon\} \leq C_1 e^{-C_2 n\epsilon^2} \quad (3.12)$$

for all $\epsilon>0$. Kiefer (1961) improved this result by showing that for each $b\in(0,2)$ there exists a constant $C_{b,m}>0$ such that for all $\epsilon>0$

$$P\{D_n \geq \epsilon\} \leq C_{b,m} e^{-(2-b)n\epsilon^2} . \quad (3.13)$$

However in none of the cited papers are explicit expressions obtained for C, C_1, C_2 and $C_{b,m}$. Among other things we will find an explicit expression for a constant C_m (depending upon m) such that for all $\epsilon > 0$

$$P\{D_n \geq \epsilon\} \leq C_m n^{2m} e^{-2n\epsilon^2}. \tag{3.14}$$

To derive (3.14) and other bounds for (3.3) we will employ techniques that were first suggested by Vapnik and Chervonenkis (1971). Their original result is the following. If $x_i \in \mathbf{R}^m$, $1 \leq i \leq n$, and if $N_{\mathcal{C}}(x_1, \ldots, x_n)$ denotes the total number of different sets in $\{(x_1, \ldots, x_n) \cap A \mid A \in \mathcal{C}\}$, then,

$$P\{\sup_{A \in \mathcal{C}} |\mu_n(A) - \mu(A)| \geq \epsilon\} \leq 4 s(\mathcal{C}, 2n) e^{-n\epsilon^2/8} \tag{3.15}$$

where

$$s(\mathcal{C}, n) = \max_{(x_1, \ldots, x_n)} N_{\mathcal{C}}(x_1, \ldots, x_n) . \tag{3.16}$$

We remark that (3.15) and related inequalities are only useful if suitable upper bounds can be found for $s(\mathcal{C}, 2n)$ for the class $\mathcal{C}$ under consideration. The second section of this chapter deals with some techniques to find such bounds. In the third section, the main results are presented.

3.2 Upper Bounds For $s(\mathcal{C}, n)$

Let $\mathcal{C}$ be a class of sets from $\mathbf{R}^m$ $(m \geq 1)$ and let $s(\mathcal{C}, n)$ be the maximal number of different sets in $\{(x_1, \ldots, x_n) \cap A \mid A \in \mathcal{C}\}$ when the maximum is taken over all $(x_1, \ldots, x_n) \in \mathbf{R}^{mn}$. It is clear that for any $\mathcal{C}$, $s(\mathcal{C}, n) \leq 2^n$. We state three lemmas. Lemma 3.1 is proved by Vapnik and Chervonenkis (1971). The proof of lemma 3.2 is trivial.

Lemma 3.1. $s(\mathcal{G},n)$ is either exactly 2^n or else upper bounded by n^b+1 where $b > 0$ is the first integer for which $s(\mathcal{G},b) < 2^b$.

Lemma 3.2. If $\mathcal{G}_1$ and $\mathcal{G}_2$ are two classes of sets from $\mathbb{R}^m$ then $s(\mathcal{G}_1\ \mathcal{G}_2,n) \le s(\mathcal{G}_1,n)\, s(\mathcal{G}_2,n)$ where $\mathcal{G}_1\mathcal{G}_2=\{A_1\cap A_2 \,|A_1\in\mathcal{G}_1, A_2\in\mathcal{G}_2\}$.

Lemma 3.3. If $\mathcal{G}$ is the class of all left half-infinite intervals from $\mathbb{R}$, then $s(\mathcal{G},n) \le 1+n$. If $\mathcal{G}$ is the class of all intervals from $\mathbb{R}$, then $s(\mathcal{G},n) \le 1+n^2$.

The proof of the first part of lemma 3.3 is trivial. For the second part, notice that $s(\mathcal{G},n) \le 1+n+(n-1)+\ldots+1 = 1+n(n+1)/2 \le 1+n^2$.

From lemmas 3.2 we can conclude the following.

Lemma 3.4. (i) If $\mathcal{G}$ is the class of all m-fold products of left half-infinite intervals from $\mathbb{R}$, then

$$s(\mathcal{G},n) \le (1+n)^m. \tag{3.17}$$

(ii) If $\mathcal{G}$ is the class of all rectangles in $\mathbb{R}^m$ (i.e., all m-fold products of intervals from $\mathbb{R}$), then

$$s(\mathcal{G},n) \le (1+n^2)^m. \tag{3.18}$$

The following lemma is also trivial.

Lemma 3.5. If $\mathcal{G}_1 \subseteq \mathcal{G}_2$ then $s(\mathcal{G}_1,n) \le s(\mathcal{G}_2,n)$.

Let $\mathcal{H}$ be the class of all closed and all open linear halfspaces in $\mathbb{R}^m$ where an open halfspace is the set of points $x=(x_1,\ldots,x_n)$ in $\mathbb{R}^m$ satisfying the inequality

$$\sum_{i=1}^{n} x_i a_i > a_0$$

for some sequence $a_0,a_1,\ldots,a_n$ of real numbers. A closed linear halfspace is just the complement of an open linear halfspace or,

equivalently, a set of points satisfying the inequality

$$\sum_{i=1}^{n} x_i a_i \geq a_0$$

for some sequence $a_0, a_1, \ldots, a_n$ of real numbers.

Let $\mathscr{S}_k, k > 0$, be the class of all open and all closed spheres in $\mathbb{R}^m$ where the norm used is

$$\|y\|_k = \begin{cases} \left(\sum_{i=1}^{m} |y_i|^k\right)^{1/k} & , k < \infty \\ \max_{1 \leq i \leq m} |y_i| & , k = \infty \end{cases} \tag{3.19}$$

and $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$. We now have

Lemma 3.6. $s(\mathscr{H}, n) \leq 1+n^{m+1}$

$s(\mathscr{S}_k, n) \leq 1+n^{2+m(k-1)}$, k even integer

$s(\mathscr{S}_\infty, n) \leq (1+n^2)^m$

$s(\mathscr{S}_1, n) \leq (1+n^2)^m$.

Proof: The first part of lemma 3.6 is shown by Vapnik and Chervonenkis (1971, p. 266). For the third part, note that $\mathscr{S}_\infty$ is contained in the class of all rectangles from $\mathbb{R}^m$ so that $s(\mathscr{S}_\infty, n) \leq (1+n^2)^m$ by lemmas 3.4 and 3.5. For the fourth part, we remark that $\mathscr{S}_1$ equals $\mathscr{S}_\infty$ after a rotation of the axes in $\mathbb{R}^m$. If $k \geq 2$, k even, then it is clear that $\mathscr{S}_k \subset \mathscr{H}^*$ where $\mathscr{H}^*$ is the class of all linear halfspaces in $\mathbb{R}^{1+m(k-1)}$. To see this, note that for $y \in \mathbb{R}^m$ and $r > 0$, a closed sphere $S(y,r)$ with center y and radius r is a linear halfspace if one considers $x_1, \ldots, x_m$,

$x_1^2,\ldots,x_m^2,\ldots,x_1^{k-1},\ldots,x_m^{k-1}$ and $\sum_{i=1}^{m} x_i^k$ as the new variables.

Therefore, by lemma 3.5 and the first part of lemma 3.6, $s(\mathcal{I}_k,n) \le s(\mathcal{K}^*,n) \le 1+n^{2+m(k-1)}$.

Q.E.D.

3.3 Main Results

Theorem 3.1. Let μ_n' and μ_n'', be the empirical measures of μ with two independent samples, one of size n and one of size n'. Let $\epsilon > 0$ and $\alpha \in (0,1)$ be such that $1-2e^{-2\alpha^2 n'\epsilon^2} > 0$. Let $\mathcal{C}$ be a class of Borel set from $\mathbb{R}^m$ such that $\sup_{A\in\mathcal{C}} |\mu_n(A)-\mu(A)|$ and $\sup_{A\in\mathcal{C}} |\mu_n'(A)-\mu_n''(A)|$ are random variables for all n,n'. Then,

$$P\{\sup_{A\in\mathcal{C}} |\mu_n(A)-\mu(A)| \ge \epsilon\} \le \frac{2s(\mathcal{C},n+n')}{1-2e^{-2\alpha^2 n'\epsilon^2}} e^{-2n\epsilon^2(1-2\alpha-2n/(n+n'))} \tag{3.20}$$

The proof of theorem 3.1 follows the argument in Vapnik and Chervonenkis (1971). We remark that α and n' are arbitrary, a freedom which can be used to obtain an estimate for $P\{D_n \ge \epsilon\}$.

Theorem 3.2. There exists a constant C_m such that for all n and $\epsilon > 0$

$$P\{D_n \ge \epsilon\} \le C_m n^{2m} e^{-2n\epsilon^2}. \tag{3.21}$$

In particular, (3.21) holds with $C_m = 4e^{3}5^{m}$.

Notice that with $\alpha=1/2, n'=n$, (3.20) implies (3.15), which is the bound of Vapnik and Chervonenkis (1971). From (3.15) and (3.17), we deduce

$$P\{D_n \geq \epsilon\} \leq 4(1+2n)^m e^{-n\epsilon^2/8} \tag{3.22}$$

which is valid for all $\epsilon > 0$ and $n \geq 1$. From lemma 3.6 and (3.15) it is also possible to derive the following inequalities.

$$P\{\sup_{A \in \mathcal{S}_k} |\mu_n(A) - \mu(A)| \geq \epsilon\} \tag{3.23}$$

$$\leq 4(1+2n)^{2+m(k-1)} e^{-n\epsilon^2/8} \quad (k \geq 2,\ k \text{ even})$$

$$P\{\sup_{A \in \mathcal{S}_k} |\mu_n(A) - \mu(A)| \geq \epsilon\} \leq 4(1+2n)^{2m} e^{-n\epsilon^2/8} . \tag{3.24}$$

We will next present a uniform counterpart to Bennett's inequality (3.5). Let μ_n' and μ_n'' be empirical measures of μ with two independent samples, both of size n. Then the following is true.

Theorem 3.3. Let $\epsilon > 0$ and let $\mathcal{C}$ be a class of Borel sets from $\mathbb{R}^m$ which is such that $\sup_{A \in \mathcal{C}} \mu_n(A)$, $\sup_{A \in \mathcal{C}} |\mu_n(A) - \mu(A)|$ and $\sup_{A \in \mathcal{C}} |\mu_n'(A) - \mu_n''(A)|$ are random variables for all n. If

$$\sup_{A \in \mathcal{C}} \mu(A) \leq g \leq 1/2 \tag{3.25}$$

then

$$P\{\sup_{A \in \mathcal{C}} |\mu_n(A) - \mu(A)| \geq \epsilon\}$$

$$\leq 4s(\mathcal{C}, 2n)\, e^{-n\epsilon^2/(64g+4\epsilon)} + 2P\{\sup_{A \in \mathcal{C}} \mu_{2n}(A) > 2g\} \tag{3.26}$$

for all n with $n \geq 8g/\epsilon^2$.

In nonparametric density estimation an important class of Borel sets is the class of all the sets with bounded radius under the norm

$\|\cdot\|_k$. The following theorem will prove very useful in upper bounding the term $P\{\sup_{A\in\mathcal{C}} \mu_{2n}(A) > 2g\}$ in (3.26).

Theorem 3.4. Let $\mathcal{C}$ be a class of Borel sets such that

$$\sup_{A\in\mathcal{C}} \sup_{y\in A, x\in A} \|y-x\| \le r < \infty \tag{3.27}$$

where $\|\cdot\|$ is any norm on $\mathbb{R}^m$. If

$$\sup_{x\in\mathbb{R}^m} \mu(S(x,2r)) \le g \le 1/2 ,$$

where $S(x,2r) \triangleq \{y: y\in\mathbb{R}^m; \|y-x\| \le 2r\}$, then

$$P\{\sup_{A\in\mathcal{C}} \mu_{2n}(A) > 2g\} \le 2ne^{-ng/10} \tag{3.28}$$

for all n with $n \ge 1/g$.

As an example, let $\mathcal{C}_r^*$ be the class of all rectangles from $\mathbb{R}^m$ with the property (3.27). Observe that

$$\sup_x \mu(S(x,2r)) \le \sup_{A\in\mathcal{C}_{2r}^*} \mu(A)$$

and that the measurability condition of theorem 3.3 is satisfied for $\mathcal{C}_r^*$ for all r. Therefore, combining lemmas 3.5, 3.6 and theorems 3.3 and 3.5, yields, for all $\epsilon > 0$,

$$P\{\sup_{A\in\mathcal{C}} |\mu_n(A) - \mu(A)| \ge \epsilon\}$$

$$\le 4(1+2n)^{2m} e^{-n\epsilon^2/(64g+4\epsilon)} + 4ne^{-ng/10} \tag{3.29}$$

for all $n \ge 1/g$, $n \ge 8g/\epsilon^2$, provided that

$$\sup_{A\in\mathcal{C}_{2r}^*} \mu(A) \le g \le 1/2 .$$

Let us make another interesting observation. When μ puts all its mass on a countable set of points, say $\{x_1, x_2, \ldots\}$, then it is clear that $\mu_n(\{x_j\}) = N_j/n$ where N_j is the number of X_i's such that $X_i = x_j$. From (3.11) we can conclude that there exists a constant $C > 0$ such that for all $\epsilon > 0$

$$P\left\{ \bigcup_{j=1}^{\infty} \{ |N_j/n - \mu(\{x_j\})| \geq \epsilon \} \right\} \leq Ce^{-2n\epsilon^2} . \tag{3.30}$$

The space from which the x_j are taken is irrelevant. We remark that (3.30) can be used to improve some of the bounds that were obtained in the proof of theorem 2.8. We should mention that (3.6) is not valid for such atomic measures μ. In fact, it is true that

$$\sup_{A \in \mathcal{B}} |\mu_n(A) - \mu(A)| \overset{n}{\to} 0 \quad \text{wp1} \tag{3.31}$$

if μ is atomic. Unfortunately, the rate of convergence to 0 is not uniform over all such μ because for all $0 < \epsilon \leq 1/4$ and all n,

$$\sup_{\substack{\text{all atomic} \\ \text{measures } \mu}} P\{ \sup_{A \in \mathcal{B}} |\mu_n(A) - \mu(A)| \geq \epsilon \} = 1 . \tag{3.32}$$

Both properties (3.31) and (3.32) are proved in the next section.

3.4 Proofs

Proof of theorem 3.1

Let $S'_n = (X_1, \ldots, X_n)$ and $S''_{n'} = (X_{n+1}, \ldots, X_{n+n'})$ where $X_1, \ldots, X_{n+n'}$ are iid random vectors from $\mathbb{R}^m$ with probability measure μ. Denote the total $(n+n')$-size sample by $S_{n+n'}$, and let μ'_n, $\mu''_{n'}$ denote the empiric probability measures for S'_n and $S''_{n'}$, respectively. For each Borel subset A of $\mathbb{R}^m$, let

$$\rho_A^{(n,n')} = |\mu_n'(A) - \mu_{n'}''(A)|$$

and let

$$\rho^{(n,n')} = \sup_{A \in \mathcal{G}} \rho_A^{(n,n')} .$$

Define

$$\pi^n = \sup_{A \in \mathcal{G}} |\mu_n'(A) - \mu(A)|$$

and note that $\rho^{(n,n')}$ and $\pi^{(n)}$ are random variables by supposition. Let

$$C_\pi = \{\pi^{(n)} > \epsilon\} \quad , \quad C_\rho = \{\rho^{(n,n')} \geq (1-\alpha)\epsilon\}$$

where $0 < \alpha < 1$ and $\epsilon > 0$. Let P, P' and P" be the probability measures induced by $S_{n+n'}$, S_n' and $S_{n'}''$, in $\mathbb{R}^{(n+n')m}$, $\mathbb{R}^{nm}$ and $\mathbb{R}^{n'm}$. We will first show that

$$P\{C_\rho\} \geq (1-2e^{-2\alpha^2 n' \epsilon^2}) P\{C_\pi\} .$$

Indeed, if $I_{C_\pi} = 1$ (i.e., $\sup_{A \in \mathcal{G}} |\mu_n'(A) - \mu(A)| > \epsilon$), then there exists an $A_0 \in \mathcal{G}$ (A_0 depends upon S_n' of course) such that $|\mu_n'(A_0) - \mu(A_0)| > \epsilon$. Thus, on $\{S_n' \in C_\pi\}$,

$$\{|\mu_{n'}''(A_0) - \mu(A_0)| \leq \alpha\epsilon\} \subseteq \{\rho_{A_0}^{(n,n')} \geq (1-\alpha)\epsilon\}$$

$$\subseteq \{\rho^{(n,n')} \geq (1-\alpha)\epsilon\} = C_\rho .$$

Therefore,

$$P\{C_\rho\} = \int_{\mathbb{R}^{(n+n')m}} I_{C_\rho} \, dP$$

$$= \int_{\mathbb{R}^{nm}} dP' \int_{\mathbb{R}^{n'm}} I_{C_\rho} dP''$$

$$\geq \int_{C_\pi} dP' \int_{\mathbb{R}^{n'm}} I_{C_\rho} dP''$$

$$\geq \int_{C_\pi} dP' \int_{\mathbb{R}^{n'm}} I_{\{|\mu''_{n'}(A_0) - \mu(A_0)| \leq \alpha\epsilon\}} dP''$$

$$\geq P'\{C_\pi\} \inf_{A_0 \in \mathcal{C}} P''\{|\mu''_{n'}(A_0) - \mu(A_0)| \leq \alpha\epsilon\}$$

$$\geq (1 - 2e^{-2\alpha^2 n' \epsilon^2}) P'\{C_\pi\}$$

by Hoeffding's inequality (3.4).

Consider $P\{C_\rho\}$. Let $T_i S_{n+n'}$ denote a permutation of the $X_1, \ldots, X_{n+n'}$ and let $\rho^{(n,n')}(i)$ and $\rho_A^{(n,n')}(i)$ be defined as $\rho^{(n,n')}$ and $\rho_A^{(n,n')}$ if $T_i S_{n+n'}$ replaces $S_{n+n'}$ in the definitions. Of course, for all integrable functions $g(S_{n+n'})$, it is true that

$$\int_{\mathbb{R}^{(n+n')m}} g(S_{n+n'}) dP = \int_{\mathbb{R}^{(n+n')m}} g(T_i S_{n+n'}) dP .$$

We say that two sets A_1 and A_2 from $\mathbb{R}^m$ are $S_{n+n'}$-equivalent if both sets define the same intersection with $\{X_1, \ldots, X_{n+n'}\}$, that is, no X_j takes values in the difference set $A_1 A_2^c \cup A_1^c A_2$, $1 \leq j \leq n+n!$. $S_{n+n'}$-equivalent sets have the following nice property. If $T_i S_{n+n'}$ is used in the definitions of μ'_n and $\mu''_{n'}$, then $\mu'_n(A_1) = \mu'_n(A_2)$ and $\mu''_{n'}(A_1) = \mu''_{n'}(A_2)$ for all possible $(n+n')!$ permutations T_i.

Given $S_{n+n'} \in \mathbb{R}^{(n+n')m}$, let $\mathcal{C}' \subset \mathcal{C}$ be a class of sets from $\mathcal{C}$ with the property that every set $A \in \mathcal{C}$, $A \notin \mathcal{C}'$ is $S_{n+n'}$-equivalent to some set $B \in \mathcal{C}'$ and that no two sets from $\mathcal{C}'$ are $S_{n+n'}$-equivalent to

each other. Obviously, regardless of $S_{n+n'}$, $\mathcal{G}'$ has never more than $s(\mathcal{G}, n+n')$ component sets. To make the dependence on $S_{n+n'}$ explicit, we will write $\mathcal{G}'(S_{n+n'})$.

Proceeding as in Vapnik and Chervonenkis (1971),

$$P\{C_\rho\} = P\{\rho^{(n,n')} \geq (1-\alpha)\epsilon\}$$

$$= \int_{\mathbb{R}^{(n+n')m}} \frac{1}{(n+n')!} \sum_{i=1}^{(n+n')!} I_{\{\rho^{(n,n')}(i) \geq (1-\alpha)\epsilon\}} dP$$

$$= \int_{\mathbb{R}^{(n+n')m}} \frac{1}{(n+n')!} \sum_{i=1}^{(n+n')!} \sup_{A \in \mathcal{G}} I_{\{\rho_A^{(n,n')}(i) \geq (1-\alpha)\epsilon\}} dP$$

$$\leq \int_{\mathbb{R}^{(n+n')m}} \frac{1}{(n+n')!} \sum_{i=1}^{(n+n')!} \sum_{A \in \mathcal{G}'(S_{n+n'})} I_{\{\rho_A^{(n,n')}(i) \geq (1-\alpha)\epsilon\}} dP$$

$$\leq \int_{\mathbb{R}^{(n+n')m}} \sum_{A \in \mathcal{G}'(S_{n+n'})} \frac{1}{(n+n')!} \sum_{i=1}^{(n+n')!} I_{\{\rho_A^{(n,n')}(i) \geq (1-\alpha)\epsilon\}} dP.$$

Let $Y_1, \ldots, Y_n, Y_{n+1}, \ldots, Y_{n+n'}$ be random variables whose values are obtained by picking at random, but without replacement, from $I_{\{X_1 \in A\}}, \ldots, I_{\{X_{n+n'} \in A\}}$ (thus, all the Y_i take values in $\{0,1\}$). Then,

$$\frac{1}{(n+n')!} \sum_{i=1}^{(n+n')!} I_{\{\rho_A^{(n,n')}(i) \geq (1-\alpha)\epsilon\}}$$

$$= P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n'}\sum_{i=n+1}^{n+n'} Y_i\right| \geq (1-\alpha)\epsilon\right\}$$

$$= P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n'}\left((n+n')\mu_{n+n'}(A) - \sum_{i=1}^{n} Y_i\right)\right| \geq (1-\alpha)\epsilon\right\}$$

$$= P\{\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu_{n+n'}(A)\right| \geq (1-\alpha)\epsilon n'/(n+n')\}$$

$$\leq 2e^{-2n((1-\alpha)\epsilon n'/(n+n'))^2}$$

$$\leq 2e^{-2n\epsilon^2(1-2\alpha-2n/(n+n'))}$$

by Hoeffding's inequality for sampling without replacement from a set of binary valued elements with mean $\mu_{n+n'}(A)$ (see lemmas 4.1 and 4.2). Notice that the last inequality holds true for all sets $A \subset \mathbb{R}^m$. So, combining the last two chains of inequalities yields

$$P\{C_\rho\} \leq \left(\int_{\mathbb{R}^{(n+n')m}} \sum_{A \in \mathcal{G}'(S_{n+n'})} dP\right) 2e^{-2n\epsilon^2(1-2\alpha-2n/(n+n'))}$$

$$\leq s(\mathcal{G}, n+n')2e^{-2n\epsilon^2(1-2\alpha-2n/(n+n'))}$$

which, together with

$$P'\{C_\pi\} \leq (P\{C_\rho\})/(1-2e^{-2\alpha^2 n'\epsilon^2})$$

concludes the proof of theorem 3.1.

Q.E.D.

Proof of theorem 3.2

In (3.20), let $\mathcal{G}$ be the class of m-fold products of left-infinite intervals from $\mathbb{R}$ and let $\alpha = 1/n\epsilon^2$, $n' = 3n^2\epsilon^2$, and let n be so large that $\alpha < 1$, that is, $1 < n\epsilon^2$. Recall that

$$s(\mathcal{G}, n+n') \leq (1+n+n')^m \leq (1+nn\epsilon^2+3n^2\epsilon^2)^m = (1+4n^2\epsilon^2)^m.$$

Then, since

$$2\alpha^2 n'\epsilon^2 = 3/2 > \ln 4$$

and

$$4\alpha^2 n\,\epsilon^2 = 2\ ,\quad 4n^2\epsilon^2/(n+n') > 4n^2\epsilon^2/(n^2\epsilon^2+3n^2\epsilon^2) = 1\ ,$$

we have that

$$P\{D_n \geq \epsilon\} \leq \frac{2(1+4n^2\epsilon^2)^m}{1-2e^{-1n4}}\, e^{-2n\epsilon^2}\, e^{4\alpha n\epsilon^2}\, e^{4n^2\epsilon^2/(n+n')}$$

$$\leq 4(1+4n^2\epsilon^2)^m\, e^{-2n\epsilon^2}\, e^3$$

which is valid for $n\epsilon^2 > 1$. If $n\epsilon^2 < 1$ however, the bound is smallest for $\epsilon = 0$ or $\epsilon = 1/\sqrt{n}$. In both cases, the bound is greater than 1, so that we can say that it is applicable for <u>all</u> n and <u>all</u> $\epsilon > 0$. Since we can assume that $\epsilon \leq 1$, we obtain

$$P\{D_n \geq \epsilon\} \leq 4e^3(1+4n^2)^m\, e^{-2n\epsilon^2} \leq 4e^3(5n^2)^m\, e^{-2n\epsilon^2}\ .$$

To see that the measurability condition of theorem 3.1 is fulfilled, note that for all $\epsilon > 0$

$$\{\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| > \epsilon\} = \{\sup_{x\in\mathbb{R}^m} |F_n(x) - F(x)| > \epsilon\} = \{\sup_{x\in D} |F_n(x) - F(x)| > \epsilon\}$$

where F_n is the empirical distribution function that corresponds to μ_n and D is a countable dense subset of $\mathbb{R}^m$. If F_n' and F_n'', correspond to μ_n' and μ_n'', respectively, then we also have that

$$\{\sup_{A\in\mathcal{A}} |\mu_n'(A) - \mu_n''(A)| > \epsilon\} = \{\sup_{x\in D} |F_n'(x) - F_n''(x)| > \epsilon\}\ .$$

Q.E.D.

<u>Proof of theorem 3.3</u>

The proof makes repeated use of Bennett's inequality (3.5). It was pointed out by Hoeffding (1963) that (3.5) also holds if

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

where $Y_1, \ldots, Y_n$ are random variables whose values are obtained by sampling without replacement from a population $y_1, \ldots, y_k$, $k \geq n$, with $y_i \in \{0,1\}$, $1 \leq i \leq k$, and

$$\mu(A) = \frac{1}{k}\sum_{i=1}^{k} y_i \ .$$

The proof of theorem 3.1 is followed with $\alpha = 1/2$, $n' = n$. Using the same notation, we first note that

$$P\{C_\rho\} \geq \inf_{A \in \mathcal{C}} P\{|\mu_n''(A) - \mu(A)| \leq \epsilon/2\}\, P\{C_\pi\}$$

$$\geq \left[1 - \sup_{A \in \mathcal{C}} \left(\frac{\mu(A)(1-\mu(A))}{n(\epsilon/2)^2}\right)\right] P\{C_\pi\}$$

$$\geq [1 - (4g/n\epsilon^2)]\, P\{C_\pi\}$$

$$\geq P\{C_\pi\}/2$$

by using Chebyshev's inequality, the fact that $n \geq 8g/\epsilon^2$ and $\sup_{A \in \mathcal{C}} \mu(A) \leq g \leq 1/2$.

As in theorem 3.1, we will upper bound $P\{C_\rho\}$. For every event E,

$$P\{C_\rho\} \leq P\{\rho^{(n,n)} \geq \epsilon/2, E\} + P\{E^c\}$$

where E^c denotes the complement of E. Let $E = \{\sup_{A \in \mathcal{C}} \mu_{2n}(A) \leq 2g\}$ which is an event by hypothesis. Thus,

$$P\{C_\rho\} \le \int_{R^{2nm}} \sum_{A\in\mathcal{G}'(S_{2n})} I_E \text{Prob}\{\left|\sum_{i=1}^{n} Y_i/n - \mu_{2n}(A)\right| \ge \epsilon/4\}\, dP$$

$$\le \int_{R^{2nm}} \sum_{A\in\mathcal{G}'(S_{2n})} 2\, e^{-n(\epsilon/4)^2/(2\mu_{2n}(A)+\epsilon/4)} I_E\, dP + P\{E^c\}$$

$$\le 2\, s(\mathcal{G},2n)\, e^{-n\epsilon^2/(64g+4\epsilon)} + P\{\sup_{A\in\mathcal{G}} \mu_{2n}(A) > 2g\}$$

since, on E, $\mu_{2n}(A) \le 2g$. Thus, for all $n \ge 8g/\epsilon^2$,

$$P\{C_\pi\} \le 4\, s(\mathcal{G},2n)\, e^{-n\epsilon^2/(64g+4\epsilon)} + 2P\{\sup_{A\in\mathcal{G}} \mu_{2n}(A) > 2g\}.$$

Q.E.D.

Proof of theorem 3.4

Let $\mathcal{G}$ satisfy (3.27) with a given $0 \le r < \infty$. Let μ^i_{2n-1} be the empirical measure for $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{2n}$ where $X_1, \ldots, X_{2n}$ are iid random vectors from R^m with probability measure μ. Note that

$$P\{\sup_{A\in\mathcal{G}} \mu_{2n}(A) > 2g\} \le P\{\bigcup_{i=1}^{2n} \{\mu_{2n}(S(X_i,2r)) > 2g\}\}$$

$$= P\{\bigcup_{i=1}^{2n} \{2n\,\mu_{2n}(S(X_i,2r)) > 2g\,2n\}\}$$

$$\le P\{\bigcup_{i=1}^{2n} \{(2n-1)\mu^i_{2n-1}(S(X_i,2r)) > 4gn-1\}\}$$

$$\le \sum_{i=1}^{2n} P\{\mu^i_{2n-1}(S(X_i,2r)) > (4gn-1)/(2n-1)\}$$

$$\le 2n\, P\{\mu^1_{2n-1}(S(X_1,2r)) > 3g/2\}$$

$$\le 2n \sup_{x\in R^m} P\{\mu^1_{2n-1}(S(x,2r)) > 3g/2\}$$

$$\le 2n \sup_{x\in R^m} P\{\mu^1_{2n-1}(S(x,2r)) - \mu(S(x,2r)) > g/2\}$$

$$\le 2n\, e^{-(2n-1)(g/2)^2/(2g+g/2)}$$

$$\le 2n\, e^{-(2n-1)g/10}$$

$$\le 2n\, e^{-ng/10}$$

for all n with ng≥1. To derive this chain of inequalities we used the one-sided version of Bennett's inequality (Bennett, 1962).

Q.E.D.

Proof of (3.31) and (3.32)

Let μ put all its mass on $\{x_1, x_2, \ldots\}$ and let $q_k = \mu(\{x_k\})$, $k=1,2,\ldots$. Let

$$U_n = \sup_{A \in \mathcal{B}} |\mu_n(A) - \mu(A)|$$

where μ_n is the empirical measure for $X_1, \ldots, X_n$. To show (3.31), we use the Borel-Cantelli lemma and the fact that for every $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P\{U_n \geq \epsilon\} < \infty .$$

Indeed, for a given $\epsilon > 0$, pick $N \geq 3$ so large that

$$\sum_{k=N+1}^{\infty} q_k < \epsilon/3$$

and note that

$$U_n = \tfrac{1}{2} \sum_{k=1}^{\infty} |\mu_n(\{x_k\}) - q_k| .$$

Therefore,

$$P\{U_n \geq \epsilon\} \leq P\{\sum_{k=1}^{N} |\mu_n(\{x_k\}) - q_k| \geq \epsilon\} + P\{\mu_n(\bigcup_{k=N+1}^{\infty} \{x_k\}) \geq 2\epsilon/3\}$$

$$\leq \sum_{k=1}^{N} P\{|\mu_n(\{x_k\}) - q_k| \geq \epsilon/N\} + P\{\mu_n(\bigcup_{k=N+1}^{\infty} \{x_k\}) - \sum_{k=N+1}^{\infty} q_k \geq \epsilon/3\}$$

$$\leq 2N e^{-2n(\epsilon/N)^2} + e^{-2n(\epsilon/3)^2} \leq (2N+1) e^{-2n\epsilon^2/N^2}$$

by Hoeffding's inequality (Hoeffding, 1963).

To show (3.32), fix $0 < \epsilon \leq \frac{1}{4}$ and $n \geq 1$. Let $q_k = 1/2n$ if $1 \leq k \leq 2n$ and 0 otherwise. Then, $P\{U_n \geq \epsilon\} \geq P\{n/2n \geq 2\epsilon\} = 1$.

Q.E.D.

Chapter 4 NONPARAMETRIC DENSITY ESTIMATION

4.1 Introduction

Let $X_1,\ldots,X_n$ be a sequence of independent, identically distributed random vectors taking values in $\mathbb{R}^m$. Assume that the common probability measure μ of the sequence is absolutely continuous with respect to Lebesgue measure with a probability density f. If μ_n is the empirical measure on $\mathcal{B}$ for $X_1,\ldots,X_n$ we can easily see that

$$\sup_{A\in\mathcal{B}} |\mu_n(A)-\mu(A)| = 1$$

since μ_n is atomic with mass $1/n$ at the points $X_1,\ldots,X_n$. Suppose now we look for an estimate μ_n of μ for which

$$\sup_{A\in\mathcal{B}} |\mu_n(A)-\mu(A)| \overset{n}{\to} 0 \text{ wpl}. \tag{4.1}$$

Assuming that μ_n has a probability density f_n we see that (4.1) will follow whenever

$$\int_{\mathbb{R}^m} |f_n(x)-f(x)|\,dx \overset{n}{\to} 0 \text{ wpl}.$$

Indeed,

$$\sup_{A\in\mathcal{B}} |\mu_n(A)-\mu(A)| = \sup_{A\in\mathcal{B}} \left| \int_A f_n(x)\,dx - \int_A f(x)\,dx \right|$$

$$\leq \sup_{A\in\mathcal{B}} \int_A |f_n(x)-f(x)|\,dx \leq \int_{\mathbb{R}^m} |f_n(x)-f(x)|\,dx . \tag{4.2}$$

This shows the importance of $\int |f_n(x)-f(x)|\,dx$ for the study of the uniform convergence properties of the corresponding measure μ_n.

Of course, from the discrimination viewpoint, we are also interested in estimates f_n for which $|f_n(x)-f(x)| \overset{n}{\to} 0$ wpl ae(μ) (see theorem 2.3). For this purpose it suffices to find an f_n for which $|f_n(x)-f(x)| \overset{n}{\to} 0$ wpl almost everywhere in x. If f is almost everywhere continuous, it suffices to establish convergence on the con-

tinuity set of f. If f is uniformly continuous, one is tempted to believe that $\sup_x |f_n(x)-f(x)| \overset{n}{\to} 0$ wpl if $|f_n(x)-f(x)| \overset{n}{\to} 0$ wpl for all x. This is not quite true but the conditions under which uniform convergence takes place are only mildly stronger than the conditions that are needed to insure the pointwise convergence for the nonparametric density estimates that will be discussed in this chapter. We will investigate the asymptotic behavior of two popular estimates.

(i) The Parzen-Rosenblatt estimate (or kernel estimate)(Parzen, 1962; Rosenblatt, 1957)

$$f_n(x) = n^{-1}\sum_{i=1}^{n} h_n^{-m} K((X_i-x)/h_n) \quad , x \in \mathbb{R}^m,$$

where $\{h_n\}$ is a sequence from $(0,\infty)$ and K is a probability density function on $\mathbb{R}^m$.

(ii) The Loftsgaarden-Quesenberry estimate (or histogram estimate, LQ estimate)(Fix and Hodges, 1951; Loftsgaarden and Quesenberry, 1965)

$$f_n(x) = C\ (k_n/n)\ /\|X_{k_n}^x - x\|^m \quad , x \in \mathbb{R}^m,$$

where $\{k_n\}$ is a sequence of integers with $1 \le k_n \le n$, C is a positive constant depending only on m and the norm $\|.\|$ and $X_{k_n}^x$ is the k_n-th nearest neighbor to x among $X_1, \ldots, X_n$.

It is clear that f_n itself is a density if the kernel method is used. With the LQ estimate, $\int f_n(x)\,dx = \infty$ for all n so that the LQ estimate is not suited for applications where one wants $\int |f_n(x)-f(x)|\,dx$ to converge to 0 in some probabilistic sense as n tends to infinity. The advantage of the LQ estimate is that it is usually easier to find a k_n for (ii) than an h_n for (i) insuring that the corresponding estimate f_n is sufficiently smooth and at the same time sufficiently

detailed. Other nonparametric density estimates not treated here include the spline method (Wahba, 1973, 1975a, 1975b) and the orthogonal series expansion. For a survey of available methods, the reader is referred to Wegman (1972).

Let us formally define the asymptotic properties that we will consider in this chapter. If B is a Borel set, then we say that $\{f_n\}$ is a pointwise weakly (strongly) consistent estimate for μ on B if there exists a density f for μ such that

$$|f_n(x) - f(x)| \xrightarrow{n} 0 \quad \text{in probability (wp1)}$$

for all $x \in B$. Further, we say that $\{f_n\}$ is a weakly (strongly) uniformly consistent estimate for μ on B if there exists a density f for μ with

$$\sup_{x \in B} |f_n(x) - f(x)| \xrightarrow{n} 0 \quad \text{in probability (wp1)} .$$

If B is omitted, it is assumed to be $\mathbb{R}^m$.

In addition to these properties we are, for kernel estimates, interested in the convergence to 0 of

$$\|f_n(z) - f(z)\|_r \triangleq \begin{cases} \left(\int |f_n(x) - f(x)|^r dx \right)^{1/r} , & 0 < r < \infty \\ \text{ess sup } |f_n(x) - f(x)| , & r = \infty \end{cases}$$

where the essential supremum is with respect to the Lebesgue measure.

4.2 Auxiliary Results

Let $Y_1, \ldots, Y_n$ be independent, identically distributed random variables with $E\{Y_1\} = 0$ and let $S_n = \sum_{i=1}^{n} Y_i$. The main tools needed below are inequalities linking $E\{|S_n/n|^k\}$ and $P\{|S_n/n| \geq \epsilon\}$ to the various statistics of Y_1.

Lemma 4.1. If Y_1 takes values in [a,b] with probability one, and if $g=b-a$, $\sigma^2 = E\{Y_1^2\}$, then, for all $\epsilon > 0$,

$$P\{|S_n/n| \geq \epsilon\} \leq 2e^{-2n(\epsilon/g)^2} \tag{4.3}$$

and

$$P\{|S_n/n| \geq \epsilon\} \leq 2e^{-n(\epsilon/g)\left((1+\sigma^2/g\epsilon)\ln(1+g\epsilon/\sigma^2) - 1\right)}$$

$$\leq 2e^{-n\epsilon^2/(2\sigma^2+g\epsilon)} . \tag{4.4}$$

We remark that (4.3) is due to Hoeffding (1963) and that (4.4) is usually attributed to Bennett (1962). The inequality on the right hand side of (4.4) is trivial if one notices that $\ln(1+u) > 2u/(2+u)$ for all $u > 0$.

Lemma 4.2. (Hoeffding, 1963). Let $k \geq n$ and let $\{y_1,\ldots,y_k\}$ be a set with $y_i \in [a,b]$ for all i. Let further $g=b-a$, and

$$k^{-1}\sum_{i=1}^{k} y_i = 0 ; \quad k^{-1}\sum_{i=1}^{k} y_i^2 = \sigma^2 .$$

If $Y_1,\ldots,Y_n$ are random variables obtained by sampling without replacement from $\{y_1,\ldots,y_k\}$, then (4.3) and (4.4) remain valid.

It should be noted that (4.3) can be strengthened for the case of sampling without replacement (Serfling, 1974).

Lemma 4.3. If Y_1 takes values in [a,b] with probability one, and if $g=b-a$, $\sigma^2 = E\{Y_1^2\}$, $E\{Y_1\} = 0$ and $r \geq 1$, then

$$E\{|S_n/n|^r\} \leq r\Gamma(r/2)(g^2/2n)^{r/2} \tag{4.5}$$

and

$$E\{|S_n/n|^r\} \leq r\Gamma(r/2)(4\sigma^2/n)^{r/2} + 2r\Gamma(r)(2g/n)^r \tag{4.6}$$

where Γ is the gamma function.

In some cases we want to obtain an upper bound for the r-th moment of S_n that is a function of $E\{|Y_1|^r\}$. We have the following result due to Whittle (1960).

Lemma 4.4. If $E\{|Y_1|^r\} < \infty$ with $r \geq 2$, then

$$E\{|S_n/n|^r\} \leq 2^{3r/2} \pi^{-1/2} \Gamma((r+1)/2) E\{|Y_1|^r\}/n^{r/2}. \tag{4.7}$$

It was shown by Rosen (1970) that (4.7) is not very tight if Y_1 has a distribution that is close to the Poisson distribution. He showed that the following lemma can sometimes provide stronger bounds.

Lemma 4.5. If $E\{|Y_1|^{2r}\} < \infty$ for some integer $r \geq 1$, and if for all $1 \leq k \leq r$,

$$E\{|Y_1|^{2k}\} \leq a^{2k} b$$

for some $a \geq 0$, $b \geq 0$, then

$$E\{|S_n/n|^{2r}\} \leq c_r \text{Max}\{(a/n)^r (ab)^r; (a/n)^{2r-1} (ab)\} \tag{4.8}$$

where c_r is a constant only depending upon r. Explicit expressions for c_r can be found in Dharmadikari and Jogdeo (1969).

4.3 Pointwise Consistency of the Parzen-Rosenblatt Estimator

Let $\{h_n\}$ be a sequence from $(0,\infty)$ and let K be a Borel measurable function from $\mathbb{R}^m$ to $[0,\infty]$. Then we call the random variable

$$f_n(x) = n^{-1} \sum_{i=1}^{n} h_n^{-m} K((X_i - x)/h_n)$$

the Parzen-Rosenblatt density estimate (Rosenblatt, 1957 ; Parzen, 1962). It is well known that $\{f_n\}$ is a pointwise weakly consistent estimate for

μ on $Q(\mu)$, the set of continuity points of μ (i.e., the set of points x for which some density of μ is continuous at x), provided that

$$h_n \overset{n}{\to} 0 , \tag{4.9}$$

$$nh_n^m \overset{n}{\to} \infty , \tag{4.10}$$

$$K \text{ is a density on } \mathbb{R}^m , \tag{4.11}$$

$$\lim_{\|x\| \to \infty} \|x\|^m K(x) = 0 , \text{ and} \tag{4.12}$$

$$\sup_{x \in \mathbb{R}^m} K(x) < \infty . \tag{4.13}$$

(For $m=1$, see Parzen (1962), Rosenblatt (1957); for $m>1$, see Cacoullos (1965)). Under much stronger conditions on $\{h_n\}$ and K (e.g., K should satisfy a local Lipschitz condition, $h_n/h_{n+1} \overset{n}{\to} 1$ and so forth) Van Ryzin (1969) showed that $\{f_n\}$ is a pointwise strongly consistent estimate for μ on $Q(\mu)$. For $m=1$, Nadaraya (1965) shows the same thing if in addition to (4.9) - (4.13), K is of bounded variation and

$$\sum_{n=1}^{\infty} e^{-\alpha n h_n^{2m}} < \infty \text{ for all } \alpha > 0 .$$

Later, Moore and Yackel (1975) mention that Nadaraya's result remains valid for $m \geq 1$.

The main result of this section (theorem 4.2) is that it suffices to add the condition

$$\sum_{n=1}^{\infty} e^{-\alpha h_n^m} < \infty \text{ for all } \alpha > 0 \tag{4.14}$$

to the list of conditions (4.9) - (4.13) in order to be able to show that $\{f_n\}$ is a pointwise strongly consistent estimate for μ on $Q(\mu)$.

Note that (4.14) is satisfied if

$$nh_n^m/\log n \overset{n}{\to} \infty \tag{4.15}$$

and this shows the closeness of the conditions (4.14) and (4.10). In fact, we will prove an inequality that is strong enough to prove both the weak and strong consistency of $\{f_n\}$.

We say that $\mu \in L_r$ $(r>0)$ if $\int f^r(x)dx < \infty$ for some density (and thus all the densities) f of μ. We say that $\mu \in L_\infty$ if ess sup $f(x) < \infty$ for some density f of μ where the essential supremum is with respect to the Lebesgue measure on $\mathbb{R}^m$. Before stating Theorem 4.1 notice that

$$E\{f_n(x)\} = E\{h_n^{-m} K((X_1-x)/h_n)\} = \int h_n^{-m} K((y-x)/h_n) f(y)dy \tag{4.16}$$

so that for any densities f and K, $E\{f_n(x)\}$ is finite almost everywhere in view of $\int (E\{f_n(x)\})dx = E\{\int f_n(x)dx\} = 1$. We first prove the following lemma which provides us with a uniform upper bound for $E\{f_n(x)\}$.

Lemma 4.6. If $\mu \in L_r$ where $1 \le r \le \infty$ and if K is an essentially bounded density, then for any density f for μ,

$$\sup_x E\{f_n(x)\} \le \|f(z)\|_r \left(\|K(z)\|_\infty / h_n^m\right)^{1/r} \tag{4.17}$$

where $\|\cdot\|_r$ is defined in section 4.1.

The following result is well-known but the proof is repeated for the sake of completeness.

Theorem 4.1. If $x \in Q(\mu)$, if K is a density on $\mathbb{R}^m$, if $h_n \overset{n}{\to} 0$ and if either $\mu \in L_\infty$ or K satisfies (4.12), then

$$|E\{f_n(x)\} - f(x)| \overset{n}{\to} 0$$

for any density f of μ that is continuous at x.

The main result of this section is the following theorem.

Theorem 4.2. If $x \in Q(\mu)$ and if (4.9)-(4.13) hold, then $|f_n(x)-f(x)| \overset{n}{\to} 0$ in probability where f is any density for μ that is continuous at x. If in addition (4.14) holds, then $|f_n(x)-f(x)| \overset{n}{\to} 0$ wp1.

If $\mu \in L_\infty$, then (4.12) is not needed in either part of the theorem.

The proof of theorem 4.2 is based upon Bennett's inequality (lemma 4.1). We remark that theorem 4.2 is essentially all we need to know about f_n to use the kernel estimate in asymptotically optimal two-step decision rules. The following theorems are rather technical. In particular, theorem 4.4 states that under the conditions of theorem 4.2, $E\{|f_n(x) - f(x)|^s\} \overset{n}{\to} 0$ for all $1 \le s \le \infty$.

Theorem 4.3.

(i) If K is a density satisfying (4.13), if $nh_n^m \overset{n}{\to} \infty$, if $\mu \in L_r$ where $1 \le r \le \infty$ and if

$$nh_n^{m(1+1/r)} \overset{n}{\to} \infty \tag{4.18}$$

then

$$\sup_x E\{|f_n(x) - E\{f_n(x)\}|^s\} \overset{n}{\to} 0$$

for all $1 \le s < \infty$.

(ii) If K is a density for which

$$\text{ess sup } K(x) < \infty \quad \text{and} \quad \text{ess sup } \|x\|^m K(x) < \infty \tag{4.19}$$

where the essential supremum is with respect to the Lebesgue measure on $\mathbb{R}^m$, and if $nh_n^m \overset{n}{\to} \infty$ and $x \in Q(\mu)$, then

$$E\{|f_n(x) - E\{f_n(x)\}|^s\} \overset{n}{\to} 0$$

for all $1 \le s < \infty$.

By Minkowski's inequality, for any f and for all s with $1 \le s < \infty$,

$$(E\{|f_n(x) - f(x)|^s\})^{1/s} \le (E\{|f_n(x) - E\{f_n(x)\}|^s\})^{1/s}$$

$$+ |E\{f_n(x)\} - f(x)| \tag{4.20}$$

so that we can combine theorem 4.3 with theorem 4.1 to get the following theorem.

Theorem 4.4. If $x \in Q(\mu)$, if $\{h_n\}$ satisfies (4.9), (4.10) and if K is a density on $\mathbb{R}^m$ satisfying (4.12) and (4.13), then

$$E\{|f_n(x) - f(x)|^s\} \overset{n}{\to} 0$$

for all s with $1 \le s < \infty$ and for all the densities f for μ that are continuous at x.

4.4 Convergence in L_r of the Parzen-Rosenblatt Estimator

We showed in the introduction why it is important that $\|f_n(z) - f(z)\|_1 \overset{n}{\to} 0$ in probability or wp1. Glick (1974) established the connection between the pointwise consistency of $\{f_n\}$ for μ and the convergence of $\|f_n(z) - f(z)\|_1$ to 0.

For $\mu \in L_2$, Nadaraya (1963) proved, under the usual conditions on K and $\{h_n\}$ and the additional condition

$$nh_n^{2m}/\log n \overset{n}{\to} \infty , \tag{4.21}$$

that $\|f_n(z) - f(z)\|_2^2 \overset{n}{\to} 0$ wp1. We will prove, among other things, that if (4.9) - (4.13) hold, and if $\mu \in L_{2r}$ and has a density f which is almost everywhere continuous on $\mathbb{R}^m$, then $E\{\|f_n(z) - f(z)\|_{2r}^{2r}\}$ tends to 0 as n tends to infinity where r is a positive integer and f is any almost everywhere continuous density for μ. Throughout this section, by almost everywhere we mean almost everywhere with respect to the

Lebesgue measure on $\mathbb{R}^m$.

Theorem 4.5. If K is a density satisfying (4.13), if $\{h_n\}$ satisfies (4.9) - (4.10) and if either (4.12) holds or $\mu \in L_\infty$, then

$$\|f_n(z) - f(z)\|_1 \xrightarrow{n} 0 \text{ in probability}$$

for any density f for μ provided that at least one density of μ is ae continuous. If in addition (4.14) holds, then $\|f_n(z)-f(z)\|_1 \xrightarrow{n} 0$ wp1.

For $r > 1$, it is of course possible that μ does not belong to L_r so that we will not be able to use Glick's theorem (Glick, 1974). Let us first observe that if $\mu \in L_r$ for some r with $1 \le r \le \infty$, then $\mu \in L_s$ for all s with $1 \le s \le r$. To see this, notice that $\mu \in L_1$ (and, in fact, $\|f(z)\|_1 = 1$ for all f) and that, by Holder's inequality,

$$\|f(z)\|_s^s = \int f^s(x)dx \le \left(\int f^r(x)dx\right)^{(s-1)/(r-1)}\left(\int f(x)dx\right)^{(r-s)/(r-1)}$$

$$= \left(\|f(z)\|_r^r\right)^{(s-1)/(r-1)} \quad (1 \le s \le r < \infty) \tag{4.22}$$

and

$$\|f(z)\|_s^s \le \|f(z)\|_1 \ \|f(z)\|_\infty^{(s-1)} \quad (1 \le s < \infty) . \tag{4.23}$$

To prove theorem 4.6, we first note that by Minkowski's inequality, for any f and r,

$$\|f_n(z)-f(z)\|_r \le \|f_n(z)-E\{f_n(z)\}\|_r + \|E\{f_n(z)\} - f(z)\|_r . \tag{4.24}$$

We have the following lemmas for both parts on the right hand side of (4.24).

Lemma 4.7. Let K be any density on $\mathbb{R}^m$, let $h_n \xrightarrow{n} 0$ and let μ have at least one ae continuous density f.

(i) If $\mu \in L_r$ with $1 \le r < \infty$, then

$$\|E\{f_n(z)\} - f(z)\|_r \xrightarrow{n} 0$$

for any ae continuous density f for μ.

(ii) If f is a uniformly continuous density for μ, then

$$\|E\{f_n(z)\} - f(z)\|_\infty \overset{n}{\to} 0 .$$

Lemma 4.8. Let K be a density on $\mathbb{R}^m$ satisfying (4.13), let $nh_n^m \overset{n}{\to} \infty$ and let r be a positive integer. If either $\mu \in L_r$ or $nh_n^{m(2-1/r)} \overset{n}{\to} \infty$, then

$$E\{\|f_n(z) - E\{f_n(z)\}\|_{2r}^{2r}\} \overset{n}{\to} 0 . \tag{4.25}$$

Lemmas 4.7, 4.8 and equation (4.24) trivially imply the following theorem.

Theorem 4.6. Let K be a density on $\mathbb{R}^m$ satisfying (4.13), let r be a positive integer, let $\{h_n\}$ satisfy (4.9) - (4.10) and let $\mu \in L_{2r}$. Then for every density f for μ, if at least one density f of μ is ae continuous,

$$E\{\|f_n(z) - f(z)\|_{2r}^{2r}\} \overset{n}{\to} 0 .$$

By Chebyshev's inequality, we have that under the conditions of theorem 4.6, for a given positive integer r,

$$\|f_n(z) - f(z)\|_{2r}^{2r} \overset{n}{\to} 0 \text{ in probability} .$$

Nadaraya (1973) proved the wpl convergence of $\|f_n(x) - f(x)\|_2^2$ under the conditions of theorem 4.6 for r=1 and under the additional requirements

$$\begin{aligned} &\text{(i)} \quad \|x\|^m K(x) \to 0 \text{ as } \|x\| \to \infty \\ &\text{(ii)} \quad nh_n^{2m}/\log n \overset{n}{\to} \infty . \end{aligned} \tag{4.26}$$

It can be shown that (4.26)(i) is not needed and that (4.26)(ii) can be relaxed to (4.14) ($nh_n^m/\log n \overset{n}{\to} \infty$) in order to be able to conclude that

$\|f_n(z) - f(z)\|_2^2 \overset{n}{\to} 0$ wpl. The proof is not given here because it follows immediately if in Nadaraya's argument, Bennett's inequality (lemma 4.1) is used instead of Prokhorov's inequality.

In the next section, we will see under what conditions $\|f_n(z) - f(z)\|_\infty \overset{n}{\to} 0$ wpl. This result can then be used together with

$$\|f_n(z) - f(z)\|_r^r \le \|f_n(z) - f(z)\|_1 \|f_n(z) - f(z)\|_\infty^{r-1} \tag{4.27}$$

and theorem 4.5 to establish, for all r with $1 \le r < \infty$, the wpl convergence of $\|f_n(z) - f(z)\|_r^r$ to 0 as n tends to infinity.

4.5 Uniform Convergence of the Parzen-Rosenblatt Estimator

The arguments that were used to prove the theorems in section 4.4 have no simple extension to L_∞. The main objective of this section is to present new techniques to prove that

$$\sup_x |f_n(x) - f(x)| \overset{n}{\to} 0 \quad \text{wpl} \tag{4.28}$$

for some density f for μ. It turns out that for a very large class of kernels in $\mathbb{R}^1$, whenever (4.28) holds for some function f, then f must be uniformly continuous (Schuster, 1969). It is therefore only natural to assume throughout this section that μ has a uniformly continuous density f. If both f and K are continuous, then it is easy to see that (4.28) is a random variable. We assume throughout that (4.28) is a random variable for all n.

Nadaraya (1965) showed that (4.28) holds in $\mathbb{R}^1$ if f is a uniformly continuous density for μ, if K is a density of bounded variation that satisfies (4.12), and if $h_n \overset{n}{\to} 0$ and $nh_n^2 / \log n \overset{n}{\to} \infty$. His argument is based on integration by parts. Foldes and Revesz (1974), also for m=1, showed that if both f and K satisfy a Lipschitz condition and $E\{\|X_1\|^\gamma\} < \infty$ for some $\gamma > 0$, then (4.28) holds under

the usual conditions on $\{h_n\}$ (i.e., (4.9), (4.14)). Using the martingale convergence theorem, Van Ryzin (1969) showed that (4.28) holds for $m \geq 1$. To the usual conditions on K and $\{h_n\}$, he adds some smoothness conditions for K and some rather restrictive conditions for $\{h_n\}$.

Among other things we will prove that (4.28) holds for $m \geq 1$ if f is a uniformly continuous density for μ, if K satisfies (4.11) - (4.13) and is Lipschitz, if $\{h_n\}$ satisfies

$$h_n \overset{n}{\to} 0 \tag{4.29}$$

and

$$nh_n^m / \log n \overset{n}{\to} \infty , \tag{4.30}$$

and if $E\{\|X_1\|^\gamma\} < \infty$ for some $\gamma > 0$. The Lipschitz condition on K and the existence condition for $E\{\|X_1\|^\gamma\}$ can be dropped and replaced by

(i) the closure of the set of discontinuities of K has Lebesgue measure 0

(ii) K has compact support.

We will assume throughout this section that the norm on $\mathbb{R}^m$ is $\|\cdot\|_\infty$. All the theorems remain valid for the norms $\|\cdot\|_k$, $k=1,2,\ldots$. These norms should not be confused with the norms of section 4.4 on the space of all densities on $\mathbb{R}^m$.

For the sake of completeness we state the following well-known result (Nadaraya, 1965; see also lemma 4.7(ii)).

Theorem 4.7. If K is a density on $\mathbb{R}^m$ and f is a uniformly continuous density for μ, then

$$\sup_x |E\{f_n(x)\} - f(x)| \overset{n}{\to} 0$$

provided that $h_n \overset{n}{\to} 0$.

Because

$$\sup_x |f_n(x)-f(x)| \le \sup_x |f_n(x)-E\{f_n(x)\}| + \sup_x |E\{f_n(x)\}-f(x)| \tag{4.31}$$

it suffices to find conditions that insure that $\sup_x |f_n(x)-E\{f_n(x)\}| \to 0$ wpl. Theorem 4.8 is an improvement of a result of Foldes and Revesz.

Theorem 4.8. If K is a density on $\mathbb{R}^m$ with

$$\text{(i)} \quad \sup_x K(x) < \infty ,$$
$$\text{(ii)} \quad \sup_x \|x\|^m K(x) < \infty , \tag{4.32}$$

and

$$\text{(iii)} \quad \sup_x |K(x+y)-K(x)| \le C\|y\| \text{ for some } C>0 \text{ and all } y \in \mathbb{R}^m,$$

and if f is a uniformly continuous density for μ with $\int \|x\|^\gamma f(x)dx < \infty$ for some $\gamma>0$, and if (4.29) and (4.30) hold, then $\{f_n\}$ is a strongly uniformly consistent estimate for μ.

The conditions that seem restrictive in theorem 4.8 are the Lipschitz condition in (4.32) and the moment condition imposed on f. With a completely different type of argument, using approximations for K and employing the bound (3.29) for deviations of empirical measures, it is possible to get rid of these conditions. Instead, let K satisfy condition (4.33) given below.

(i) K is a probability density on $\mathbb{R}^m$;
(ii) $\sup_x K(x) < \infty$; (4.33)
(iii) K has compact support, i.e. there exists a $\rho>0$ such that $\int_{[-\rho,+\rho]^m} K(x)dx = 1$;
(iv) the closure of the set of discontinuities of K has Lebesgue measure 0.

We remark that (4.33)(iv) is not a very restrictive condition at all. We prove the following theorem.

Theorem 4.9. If K satisfies (4.33), if (4.29) and (4.30) hold and if f is a uniformly continuous density for μ, then $\{f_n\}$ is a strongly uniformly consistent estimate for μ.

One of the conditions in (4.33) is that K should have a compact support. One may wonder what happens if K does not have a compact support, for example, if K is gaussian. In the next theorem we will give a partial answer to this question. We will find that the conditions to be imposed on $\{h_n\}$ depend upon the rate of decrease to 0 of the tail of K. Consider the following condition to be used instead of (4.33)(iii).

There exists a continuous and monotonically decreasing function $u:[0,\infty) \to [0,\infty)$ with

$$\text{(i)} \quad \int_0^\infty z^{m-1} u(z)\,dz < \infty \quad \text{and}$$
$$\text{(ii)} \quad K(x) \le u(\|x\|) \text{ , all } x \in \mathbb{R}^m. \tag{4.34}$$

The condition (4.34) states that K must be dominated by a bell-shaped and integrable function u. The condition (4.34)(i) implies that

$$\int_{\mathbb{R}^m} u(\|x\|)\,dx < \infty$$

and, by the monotonicity of u, it is not hard to show that $\|x\|^m u(\|x\|) \to 0$ as $\|x\| \to \infty$. Hence, $\|x\|^m K(x) \to 0$ as $\|x\| \to \infty$, in which we recognize the classical regularity condition for K (see (4.12)). Although $z^m u(z) \to 0$ as $z \to \infty$ does not imply (4.34)(i), it is a very close condition indeed. The following functions of z satisfy the requirements imposed on u :

$$1/(1+z)^{m(1+\beta)} \text{ and } 1/((1+z)^m(1+\log(1+z))^{1+\beta}) \text{ , } \beta>0.$$

Needless to say, if K satisfies (4.34), then it is possible to find a continuous inverse u^{-1} defined on some compact set $[0,a]$ with the properties that u^{-1} is monotonically decreasing and $u^{-1}(z) \to \infty$ as $z \to 0$. Notice further that $u^{-1}(z) \leq (a_0/z)^{1/m}$ for some a_0 with $0 < a_0 < \infty$. We have the following theorem.

Theorem 4.10. If K satisfies (4.33) where (4.33)(iii) is replaced by (4.34), if (4.29) and (4.30) hold, if f is a uniformly continuous density for μ and if, in addition, for all $\epsilon > 0$,

$$nh_n^m / (u^{-1}(\epsilon h_n^m))^m \log n \overset{n}{\to} \infty \tag{4.35}$$

then $\{f_n\}$ is a strongly uniformly consistent estimate for μ.

Remark. Since $u^{-1}(z) \leq (a_0/z)^{1/m}$ for some $a_0 < \infty$, it is clear that (4.35) is always fulfilled if

$$nh_n^{2m} / \log n \overset{n}{\to} \infty. \tag{4.36}$$

However, depending upon the rate of decrease of K to 0 as $\|x\| \to \infty$, weaker conditions than (4.36) can be obtained. For example, if $K(x) \leq a_0 / \|x\|^{\alpha m}$ where $\alpha \geq 1$, then the condition

$$nh_n^{m(1+1/\alpha)} / \log n \overset{n}{\to} \infty$$

is sufficient for (4.35). Theorem 4.9 can be viewed as an extreme case where $\alpha = \infty$.

4.6 The Loftsgaarden-Quesenberry Estimator

In this section we study the asymptotic properties of a generalized version of the LQ estimate (Fix and Hodges, 1951; Loftsgaarden and Quesenberry, 1965). Let $\{k_{1n}\}$ and $\{k_{2n}\}$ be two sequences of positive integers such that

$$k_{1n} \overset{n}{\to} \infty, \tag{4.37}$$

$$1 \le k_{1n} \le k_{2n} \le n \quad , \text{ all } n \, , \tag{4.38}$$

and

$$k_{2n}/n \overset{n}{\to} 0 \, , \tag{4.39}$$

and consider the estimate

$$f_n(x) = \sum_{j=k_{1n}}^{k_{2n}} \alpha_{jn}(j/n)(2\|X_j^x - x\|)^{-m} \quad , \; x \in \mathbb{R}^m, \tag{4.40}$$

where

(i) $\alpha_n = (\alpha_{k_{1n}n}, \ldots, \alpha_{k_{2n}n})$ is a probability distribution.

(ii) $\|\cdot\|$ is the $\|\cdot\|_\infty$ norm. All the results of this section remain valid for standard norms such as $\|\cdot\|_k$, $k=1,2,\ldots$ provided that in (4.40) $(2\|X_j^x - x\|)^{+m}$ is replaced by the volume of the sphere with radius $\|X_j^x - x\|$.

(iii) X_j^x is the j-th nearest neighbor to x among $X_1, \ldots, X_n$ where ties are broken at random. Thus, $\|X_1^x - x\| \le \ldots \le \|X_n^x - x\|$.

Notice that $\sup_x |f_n(x) - f(x)|$ is a random variable if f is continuous since it is possible to replace the supremum over $\mathbb{R}^m$ by the supremum over a countable dense subset of $\mathbb{R}^m$ in view of the continuity of $|f_n - f|$. We remark that (4.40) is a "smooth" version of the original LQ estimate (Loftsgaarden and Quesenberry, 1965)

$$f_n(x) = (k_n/n)(2\|X_{k_n}^x - x\|)^{-m} \quad , \; x \in \mathbb{R}^m, \tag{4.41}$$

where (4.41) is obtained from (4.40) by letting $k_{1n}=k_{2n}=k_n$. In the original paper of Loftsgaarden and Quesenberry, the norm was $\|\cdot\|_2$ and the coefficient in (4.41) was $((k_n-1)/n)$ but both changes are

irrelevant to the results that are presented below. Loftsgaarden and Quesenberry (1965) showed that $f_n(x) \xrightarrow{n} f(x)$ in probability for each x at which f is continuous and positive, provided that $k_n/n \xrightarrow{n} 0$ and $k_n \xrightarrow{n} \infty$. Wagner (1973) extended this result to convergence wp1 under the additional condition (4.42)

$$\sum_{n=1}^{\infty} e^{-\alpha k_n} < \infty \text{ for all } \alpha > 0 . \tag{4.42}$$

For m=1, Moore and Henrichon (1969) showed that $\sup_x |f_n(x) - f(x)| \xrightarrow{n} 0$ in probability if f is uniformly continuous and positive on $(-\infty,+\infty)$, if $k_n/n \xrightarrow{n} 0$ and if

$$k_n/\log n \xrightarrow{n} \infty . \tag{4.43}$$

Kim and Van Ryzin (1975), also for m=1, prove the same thing for a slightly different type of estimate under essentially the same conditions. We remark, once and for all, that (4.43) implies (4.42) and that (4.42) is sufficient for (4.43) if $\{k_n\}$ is nondecreasing.

We will prove the strong uniform consistency of $\{f_n\}$ (defined by (4.40) for n=1,2,...) for μ on $\mathbb{R}^m$ (where $m \geq 1$) if there exists a uniformly continuous density f for μ, if (4.37) - (4.39) hold and if

$$k_{1n}^2/k_{2n} \log n \xrightarrow{n} \infty . \tag{4.44}$$

For $k_{1n}=k_{2n}$, the condition (4.44) reduces to (4.43). To prove the uniform consistency of $\{f_n\}$, we use some of the uniform inequalities for the deviation of empirical measures that were proved in chapter 3. We first prove a pointwise consistency theorem for the estimate (4.40).

Theorem 4.11. Let $\{k_{1n}\}$ and $\{k_{2n}\}$ be positive integer sequences satisfying (4.37)-(4.39) andl let $\{f_n\}$ be defined by (4.40) where $\{\alpha_n\}$ is an arbitrary sequence of probability vectors satisfying (4.40)(i). Let $x \in Q(\mu)$ and let f be a density for μ that is continuous at x.

(i) If $\sup_n (k_{2n}-k_{1n}) < \infty$, then

$$|f_n(x)-f(x)| \xrightarrow{n} 0 \text{ in probability.} \tag{4.45}$$

(ii) If, in addition,

$$\sum_{n=1}^{\infty} e^{-\alpha k_{1n}^2/k_{2n}} < \infty \quad \text{for all } \alpha>0, \tag{4.46}$$

then

$$|f_n(x)-f(x)| \xrightarrow{n} 0 \text{ wpl.} \tag{4.47}$$

(iii) If (4.44) holds, then (4.47) is valid.

The pointwise consistency theorems of Loftsgaarden and Quesenberry (1965) and Wagner (1973) can be derived from theorem 4.11 by letting $k_{1n}=k_{2n}=k_n$. The main result of this section is the following theorem.

Theorem 4.12. Let $\{k_{1n}\}$ and $\{k_{2n}\}$ be positive integer sequences satisfying (4.37)-(4.39) and let $\{f_n\}$ be defined by (4.40) where $\{\alpha_n\}$ is an arbitrary sequence of probability vectors satisfying (4.40)(i). If f is a uniformly continuous density for μ and if (4.44) holds, then

$$\sup_x |f_n(x)-f(x)| \xrightarrow{n} 0 \text{ wpl.}$$

4.7 Proofs

Proof of lemma 4.3

We will make repeated use of the gamma function identity

$$\Gamma(\alpha)\beta^{\alpha} = \int_0^{\infty} y^{\alpha-1} e^{-y/\beta}\, dy \ , \ \alpha, \beta > 0 \ .$$

Recall that for α integer, $\Gamma(\alpha) = (\alpha-1)!$

By Hoeffding's inequality (4.3),

$$E\{|S_n/n|^r\} = \int_0^{\infty} P\{|S_n/n|^r > u\}\, du$$

$$= \int_0^{\infty} \frac{r}{2} u^{\frac{r}{2}-1} P\{|S_n/n| > \sqrt{u}\}\, du$$

$$\leq 2\int_0^{\infty} \frac{r}{2} u^{\frac{r}{2}-1} e^{-2nu/g^2}\, du$$

$$= r\Gamma(r/2)(g^2/2n)^{r/2} \ .$$

By Bennett's inequality (4.4),

$$E\{|S_n/n|^r\} = \int_0^{\infty} r\, u^{r-1} P\{|S_n/n| > u\}\, du$$

$$\leq 2\int_0^{2\sigma^2/g} r u^{r-1} e^{-nu^2/4\sigma^2}\, du + \int_{2\sigma^2/g}^{\infty} r u^{r-1} e^{-nu/2g}\, du$$

$$\leq 2\int_0^\infty \frac{r}{2} u^{\frac{r}{2}-1} e^{-nu/4\sigma^2}\, du + 2\int_0^\infty r u^{r-1} e^{-nu/2g}\, du$$

$$= 2\frac{r}{2}\Gamma\left(\frac{r}{2}\right)\left(\frac{4\sigma^2}{n}\right)^{r/2} + 2r\Gamma(r)\left(\frac{2g}{n}\right)^{r}.$$

Q.E.D.

Proof of lemma 4.6

Let $1 \leq r \leq \infty$, let f be a density for μ and let $\mu \in L_r$. Then, with $\frac{1}{r} + \frac{1}{s} = 1$, we have for all $x \in \mathbb{R}^m$ by twice applying Hölder's inequality,

$$E\{f_n(x)\} = \int h_n^{-m} K((y-x)/h_n)\, f(y)\, dy$$

$$\leq \|f(z)\|_r \, \|h_n^{-m} K(z/h_n)\|_s$$

$$\leq \|f(z)\|_r \, \|h_n^{-m} K(z/h_n)\|_1^{1/s} \, \|h_n^{-m} K(z/h_n)\|_\infty^{(s-1)/s}$$

$$\leq \|f(z)\|_r \, (\|K(z)\|_\infty / h_n^m)^{1-1/s}$$

$$= \|f(z)\|_r \, (\|K(z)\|_\infty / h_n^m)^{1/r}.$$

Q.E.D.

Proof of theorem 4.1

Let $x \in Q(\mu)$ and let f be a density that is continuous at x. Find $\delta > 0$ so small that $|f(y+x) - f(x)| < \epsilon/2$ for all $y \in \mathbb{R}^m$ with $\|y\| < \delta$. Then,

$$|E\{f_n(x)\}| \leq \int |f(x+y) - f(x)|\, h_n^{-m} K(y/h_n)\, dy$$

$$< \epsilon/2 + f(x) \int_{\{y:\, \|y\| \geq \delta\}} h_n^{-m} K(y/h_n)\, dy$$

$$+ \int_{\{y:\ \|y\| \geq \delta\}} h_n^{-m} K(y/h_n)\, f(x+y)\, dy \quad .$$

If $\mu \in L_\infty$, then the two last terms are upper bounded by

$$(f(x) + \|f(z)\|_\infty) \int_{\{y:\|y\| \geq \delta/h_n\}} K(y)\, dy$$

which tends to 0 as $n \to \infty$ if $h_n \overset{n}{\to} 0$ and K is integrable. If $\mu \notin L_\infty$ but $\|x\|^m K(x) \to 0$ as $\|x\| \to \infty$, then we need only consider

$$\int_{\{y:\ \|y\| \geq \delta\}} h_n^{-m} K(y/h_n)\, f(x+y)\, dy$$

$$\leq \int_{\{y:\ \|y\| \geq \delta\}} \|y\|^{-m} (\|y\|/h_n)^m K(y/h_n)\, f(x+y)\, dy$$

$$\leq \delta^{-m} \sup_{\{y:\|y\| \geq \delta/h_n\}} \|y\|^m K(y) \overset{n}{\to} 0 \ .$$

Q.E.D.

Proof of theorem 4.2

Let f be a density for μ that is continuous at x. From theorem 4.1 and

$$|f_n(x) - f(x)| \leq |f_n(x) - E\{f_n(x)\}| + |E\{f_n(x)\} - f(x)|$$

it suffices to show that $|f_n(x) - E\{f_n(x)\}| \overset{n}{\to} 0$ either in probability or wpl under the conditions of the theorem.

Notice that $f_n(x) - E\{f_n(x)\} = \left(\sum_{i=1}^{n} Y_i\right)/n$ where $Y_1, \ldots, Y_n$ are iid random variables with $E\{Y_1\} = 0$, $|Y_1| \leq h_n^{-m} \|K(z)\|_\infty$ wpl and

$$E\{Y_1^2\} \leq \int (h_n^{-m} K((y-x)/h_n))^2 f(y)\, dy$$

$$\leq h_n^{-m} \|K(z)\|_\infty \int h_n^{-m} K(y/h_n)\, f(y+x)\, dy = \|K(z)\|_\infty h_n^{-m} E\{f_n(x)\}.$$

In the proof of theorem 4.1, we have seen that if $\mu \in L_\infty$ or $\|x\|^m K(x) \to 0$ as $\|x\| \to \infty$, and $h_n \overset{n}{\to} 0$, then $E\{f_n(x)\} \overset{n}{\to} f(x)$. Thus there exists a constant $C_0 > 0$ depending upon x such that for all n, $E\{Y_1^2\} \leq$

$C_0\|K(z)\|_\infty/h_n^m$. By (4.4), with $\sigma^2=C_0\|K(z)\|_\infty/h_n^m$ and $g=2\|K(z)\|_\infty/h_n^m$, we have

$$P\{|f_n(x)-E\{f_n(x)\}|\geq\epsilon\} \leq 2e^{-n\epsilon^2/(2C_0+2\epsilon)(\|K(z)\|_\infty h_n^{-m})}$$
$$= 2e^{-C_1 n h_n^m}$$

where $C_1=\epsilon^2/(2(C_0+\epsilon)\|K(z)\|_\infty)$. The "in probability" part of the theorem follows from the arbitrariness of ϵ and (4.10). For the wpl part, notice that (4.14) implies that

$$\sum_{n=1}^{\infty} P\{|f_n(x)-E\{f_n(x)\}|\geq\epsilon\} < \infty$$

for all $\epsilon>0$. The Borel-Cantelli lemma then implies the second part of theorem 4.2.

Q.E.D.

Proof of theorem 4.3

To show (ii), let $1\leq s<\infty$ and let x be a continuity point of f. Recall from the proof of theorem 4.2 that

$$f_n(x)-E\{f_n(x)\} = (\sum_{i=1}^{n} Y_i)/n$$

where the Y_i are iid random variables with zero mean, ess sup Y_1 - ess inf $Y_1 \leq 2\|K(z)\|_\infty/h_n^m$, and

$$E\{Y_1^2\} \leq \|K(z)\|_\infty h_n^{-m} E\{f_n(x)\} \leq C_0\|K(z)\|_\infty/h_n^m.$$

By lemma 4.3,

$$E\{|f_n(x)-E\{f_n(x)\}|^s\} \leq s\Gamma(s/2)(4C_0\|K(z)\|_\infty/nh_n^m)^{s/2}$$
$$+ 2s\Gamma(s)(4\|K(z)\|_\infty/nh_n^m)^s \overset{n}{\to} 0$$

in view of (4.10) and $\|K(z)\|_\infty<\infty$.

To show (i), let $1\leq s<\infty$ and let $\mu\in L_r$ $(1\leq r\leq\infty)$. By lemma 4.6,

$$E\{Y_1^2\} \leq \|K(z)\|_\infty h_n^{-m} E\{f_n(x)\}$$
$$\leq \|K(z)\|_\infty \|f(z)\|_r h_n^{-m}(\|K(z)\|_\infty/h_n^m)^{1/r}$$

and thus, by lemma 4.3,

$$E\{|f_n(x)-E\{f_n(x)\}|^s\} \leq 2s\,\Gamma(s)\,(4\|K(z)\|_\infty/n h_n^m)^s + s\,\Gamma(s/2)\left(4\|K(z)\|_\infty^{1+1/r}\|f(z)\|_r/n h_n^{m(1+1/r)}\right)^{s/2}$$

which tends to 0 as $n\to\infty$ in view of (4.10), (4.18) and $\mu\in L_r$.

Q.E.D.

Proof of theorem 4.5

Theorem 4.5 is a corollary of theorem 4.2 and a theorem of Glick (1974), which states that if f_n satisfies some measurability conditions (that are satisfied here) and f_n is a density for all n, and $f_n(x) \xrightarrow{n} f(x)$ in probability (wp1) ae (Lebesgue measure on $\mathbb{R}^m$), then

$$\int |f_n(x)-f(x)|\,dx \xrightarrow{n} 0 \text{ in probability (wp1).}$$

Q.E.D.

Proof of lemma 4.7

Let f be an ae continuous density for μ and let $\mu\in L_r$ where $1\leq r<\infty$. Note that

$$\|E\{f_n(z)\} - f(z)\|_r^r = \int\left|\int (f(y+x)-f(x))h_n^{-m}K(y/h_n)\,dy\right|^r dx$$

$$\leq \iint |f(y+x)-f(x)|^r h_n^{-m}K(y/h_n)\,dy\,dx$$

$$= \int h_n^{-m}K(y/h_n)\left(\int |f(x+y)-f(x)|^r dx\right) dy$$

by Jensen's inequality and Tonelli's theorem. But

$$\lim_{\|y\|\to 0} |f(x+y)-f(x)|^r = 0 \quad \text{ae}$$

by the ae continuity of f. Since, by the c_r-inequality,

$$|f(x+y)-f(x)|^r \leq 2^{r-1}(f^r(x+y) + f^r(x))$$

and

$$2^{r-1}\int(f^r(x+y)+f^r(x))\,dx = 2^r\|f(z)\|_r^r < \infty,$$

we have by a version of the Lebesgue dominated convergence theorem that

$$\int|f(x+y)-f(x)|^r\,dx \to 0 \quad \text{as } \|y\|\to 0.$$

Find δ small enough so that

$$\int|f(x+y)-f(x)|^r\,dx \le \epsilon/2$$

for all $\|y\|\le\delta$. Then,

$$\begin{aligned}\|E\{f_n(z)\}-f(z)\|_r^r &\le \epsilon/2 + \left(\int_{\{y:\|y\|>\delta\}} h_n^{-m}K(y/h_n)\,dy\right)2^r\|f(z)\|_r^r \\ &= \epsilon/2 + 2^r\|f(z)\|_r^r \int_{\{y:\|y\|>\delta/h_n\}} K(y)\,dy \\ &< \epsilon\end{aligned}$$

for all n large enough since $\mu\in L_r$, K is a density and $h_n\to 0$. The first part of lemma 4.7 follows from the arbitrariness of ϵ.

For the second part of lemma 4.7, let f be a uniformly continuous density for μ. Then,

$$\begin{aligned}\|E\{f_n(z)\}-f(z)\|_\infty &\le \sup_x |E\{f_n(x)\}-f(x)| \\ &\le \sup_x \int|f(y+x)-f(x)|\,h_n^{-m}K(y/h_n)\,dy \\ &\le \int_{\{y:\|y\|<\delta\}}(\epsilon/2)h_n^{-m}K(y/h_n)\,dy + \int_{\{y:\|y\|\ge\delta/h_n\}}\|f(z)\|_\infty K(y)dy\end{aligned}$$

where $\delta>0$ is so small that $\sup_x|f(y+x)-f(x)| < \epsilon/2$ for all y with $\|y\|<\delta$. Thus, since K is a density, $\mu\in L_\infty$ and $h_n\to 0$, we have that

$$\|E\{f_n(z)\}-f(z)\|_\infty < \epsilon$$

for all n large enough. The second part of lemma 4.7 follows from the arbitrariness of ϵ. Q.E.D.

Proof of lemma 4.8

Let r be a positive integer. Note that by Jensen's inequality and by Tonelli's theorem,

$$\left(E\{\|f_n(z)-E\{f_n(z)\}\|_{2r}\}\right)^{2r} \le E\{\|f_n(z)-E\{f_n(z)\}\|_{2r}^{2r}\}$$

$$= \int E\{|f_n(x)-E\{f_n(x)\}|^{2r}\}\,dx .$$

Now, $f_n(x)-E\{f_n(x)\} = (\sum_{i=1}^{n} Y_i)/n$ where the Y_i are iid random variables with $E\{Y_1\}=0$. Furthermore, for all k with $1 \le k \le r$,

$$E\{|Y_1|^{2k}\} = E\{|h_n^{-m}\left(K((X_1-x)/h_n)-E\{K((X_1-x)/h_n)\}\right)|^{2k}\}$$

$$\le 2^{2k-1}\left(E\{|h_n^{-m}K((X_1-x)/h_n)|^{2k}\} + |E\{h_n^{-m}K((X_1-x)/h_n)\}|^{2k}\right)$$

$$\le 2^{2k}\, E\{|h_n^{-m}K((X_1-x)/h_n)|^{2k}\}$$

$$\le 2^{2k}\,(h_n^{-m}\|K(z)\|_\infty)^{2k-1} E\{h_n^{-m}K((X_1-x)/h_n)\} .$$

By Rosen's theorem (lemma 4.5) there exists a universal constant c_r only depending upon r such that

$$E\{|f_n(x)-E\{f_n(x)\}|^{2r}\} \le c_r \operatorname{Max}\Big((2\|K(z)\|_\infty/n h_n^m)^r (2E\{f_n(x)\})^r;$$

$$(2\|K(z)\|_\infty/n h_n^m)^{2r-1}(2E\{f_n(x)\})\Big).$$

If $\mu \in L_r$, then we have by Tonelli's theorem,

$$E\{\|f_n(x)-E\{f_n(x)\}\|_{2r}^{2r}\} \le c_r(2\|K(z)\|_\infty/n h_n^m)^r 2^r \int (E\{f_n(x)\})^r dx$$

$$+ 2c_r(2\|K(z)\|_\infty/n h_n^m)^{2r-1} \xrightarrow{n} 0$$

in view of (4.10), (4.13) and

$$\int (E\{f_n(x)\})^r dx \le \iint h_n^{-m}K((y-x)/h_n)\, f^r(y)\,dy\,dx$$

$$= \int f^r(y)\,dy = \|f(z)\|_r^r < \infty .$$

If $\mu \notin L_r$ but $n h_n^{m(2-1/r)} \overset{n}{\to} \infty$, then we have by Tonelli's theorem,

$$E\{\|f_n(z)-E\{f_n(z)\}\|_{2r}^{2r}\} \le c_r(2\|K(z)\|_\infty/n h_n^m)^r 2^r(\|K(z)\|_\infty/h_n^m)^{r-1}$$

$$+ 2c_r(2\|K(z)\|_\infty/n h_n^m)^{2r-1} \overset{n}{\to} 0.$$

Q.E.D.

Proof of theorem 4.7

Let $\epsilon>0$ be arbitrary and pick $\delta>0$ so small that $|f(x+y)-f(x)|<\epsilon/2$ whenever $\|y\|<\delta$. Then,

$$\sup_x |E\{f_n(x)\}-f(x)| \le \sup_x \int |f(x+y)-f(x)| h_n^{-m} K(y/h_n)\, dy$$

$$\le \epsilon/2 + \int_{\{y:\|y\|\ge\delta/h_n\}} (\sup_x f(x))\, K(y)\, dy < \epsilon$$

for all n large enough in view of $\sup_x f(x) < \infty$, $\int K(y)\,dy < \infty$ and $h_n \overset{n}{\to} 0$.

Q.E.D.

Proof of theorem 4.8

Let $S(x,r)=\{y : y\in\mathbb{R}^m, \|y-x\| \le r\}$ where $x\in\mathbb{R}^m$ and $r\ge 0$. Let further $K_1=\sup_x K(x)$, $K_2=\sup_x \|x\|^m K(x)$, $K_3=\sup_x f(x)$ and $K_4=2^\gamma \int \|x\|^\gamma f(x)\,dx$. Let $\epsilon>0$ be arbitrary. Define the positive number sequences $\{a_n\}$ and $\{b_n\}$ and the integer sequence $\{k_n\}$ as follows. Let

$$a_n = (16 K_1 K_4/\epsilon h_n^m)^{1/\gamma},$$

$$b_n = \epsilon h_n^{m+1}/8C,$$

and

$$k_n = [2a_n/b_n]$$

where C is the constant of (4.32)(iii) and [u] stands for the smallest integer not smaller than u.

First find $(k_n)^m$ points $y_i, i=1,\ldots,k_n^m$, all belonging to

$S(0,a_n)$ such that for all x in $S(0,a_n)$ there exists a y_i with $\|y_i-x\|<b_n$. Clearly,

$$P\{\sup_x |f_n(x)-E\{f_n(x)\}| \geq \epsilon\} \leq P\{\sup_{x:\|x\|>a_n} |f_n(x)-E\{f_n(x)\}| \geq \epsilon\}$$

$$+ P\{\sup_{x:\|x\|\leq a_n} |f_n(x)-E\{f_n(x)\}| \geq \epsilon\}$$

and

$$P\{\sup_{x:\|x\|\leq a_n} |f_n(x)-E\{f_n(x)\}| \geq \epsilon\}$$

$$\leq \sum_{j=1}^{k_n} P\{\sup_{x\in S(y_j,b_n)} |f_n(x)-E\{f_n(x)\}| \geq \epsilon\}$$

$$\leq k_n^m \sup_{1\leq j\leq k_n^m} P\{\sup_{x\in S(y_j,b_n)} |f_n(x)-E\{f_n(x)\}| \geq \epsilon\}$$

$$\leq k_n^m \sup_{1\leq j\leq k_n^m} \Big(P\{\sup_{x\in S(y_j,b_n)} |f_n(x)-f_n(y_j)| \geq \epsilon/3\}$$

$$+ P\{\sup_{x\in S(y_j,b_n)} |E\{f_n(x)\}-E\{f_n(y_j)\}| \geq \epsilon/3\}$$

$$+ P\Big\{\Big|\sum_{i=1}^{n} (K((X_i-y_j)/h_n) - E\{K((X_i-y_j)/h_n)\})/nh_n^m\Big| \geq \epsilon/3\Big\}\Big).$$

Note that

$$\sup_{x\in S(y_j,b_n)} |f_n(x)-f_n(y_j)|$$

$$\leq h_n^{-m} \sup_{(x,z)\in S(y_j,b_n)x\mathbb{R}^m} |K((z-x)/h_n) - K((z-y_j)/h_n)|$$

$$\leq C\, b_n/h_n^{m+1} = \epsilon/8 < \epsilon/3$$

and that

$$\sup_{x\in S(y_j,b_n)} |E\{f_n(x)\}-E\{f_n(y_j)\}| \leq E\{\sup_{x\in S(y_j,b_n)} |f_n(x)-f_n(y_j)|\}$$

$$< \epsilon/3 .$$

We thus have that

$$P\{\sup_{x:\|x\|\le a_n} |f_n(x)-E\{f_n(x)\}| \ge \epsilon\}$$
$$\le k_n^m \sup_{1\le j\le k_n^m} P\{\left|\sum_{i=1}^{n}\left(K((X_i-y_j)/h_n)-E\{K((X_i-y_j)/h_n)\}\right)/nh_n^m\right| \ge \epsilon/3\}.$$

Consider the iid random variables $Y_1,\ldots,Y_n$ where

$$Y_i = \left(K((X_i-y_j)/h_n)-E\{K((X_i-y_j)/h_n)\}\right)/h_n^m \quad, 1\le i\le n,$$

and note that $E\{Y_1\}=0$, $E\{Y_1^2\}\le K_1K_3/h_n^m$, and ess sup Y_1 - ess inf Y_1 $\le K_1/h_n^m$. Thus, by Bennett's inequality and $k_n \le 1+2a_n/b_n$, we have,

$$P\{\sup_{x:\|x\|\le a_n} |f_n(x)-E\{f_n(x)\}| \ge \epsilon\}$$
$$\le (1+2a_n/b_n)^m 2\, e^{-C_1 n h_n^m}$$

where

$$C_1 = (\epsilon/3)^2/(2K_1K_3+K_1\epsilon/3).$$

Next,

$$P\{\sup_{x:\|x\|>a_n} |f_n(x)-E\{f_n(x)\}| \ge \epsilon\}$$
$$\le P\{\sup_{x:\|x\|>a_n} |f_n(x)| \ge \epsilon/2\} + P\{\sup_{x:\|x\|>a_n} f(x) \ge \epsilon/4\}$$
$$+ P\{\sup_x |f(x)-E\{f_n(x)\}| \ge \epsilon/4\}.$$

The last probability on the right-hand side is 0 for all n large enough by theorem 4.7. The second probability on the right-hand side is 0 for all n large enough since the uniform continuity of f and the integrability of f imply that $f(x)\to 0$ as $\|x\|\to\infty$, and since $a_n \overset{n}{\to} \infty$. Further,

$$P\{\sup_{x:\|x\|>a_n} f_n(x) \geq \epsilon/2\}$$

$$\leq P\{\sup_{x:\|x\|>a_n} (1/nh_n^m) \sum_{i:\|X_i-x\|>a_n/2} K((X_i-x)/h_n) \geq \epsilon/4\}$$

$$+ P\{\sup_{x:\|x\|>a_n} (1/nh_n^m) \sum_{i:\|X_i-x\|\leq a_n/2} K((X_i-x)/h_n) \geq \epsilon/4\}.$$

$$\leq P\{(1/nh_n^m)\, nh_n^m\, K_2/(a_n/2)^m \geq \epsilon/4\}$$

$$+ P\{(1/nh_n^m) \sum_{i:\|X_i\|\geq a_n/2} K_1 \geq \epsilon/4\}.$$

The first probability is 0 for all n large enough since $a_n \to \infty$. Further,

$$P\{\|X_1\| \geq a_n/2\} \leq E\{\|X_1\|^\gamma\}/(a_n/2)^\gamma = K_4/a_n^\gamma = \epsilon h_n^m/16K_1$$

so that $P\{(1/nh_n^m) \sum_{i:\|X_i\|\geq a_n/2} K_1 \geq \epsilon/4\}$

$$\leq P\{\sum_{i=1}^{n} \left(I_{\{\|X_i\|\geq a_n/2\}} - P\{\|X_i\| \geq a_n/2\}\right)/n \geq \epsilon h_n^m/8K_1\}$$

$$\leq e^{-C_2 n h_n^m}$$

by Bennett's inequality where $C_2 = \epsilon/16K_1$. Thus, combining bounds, we have for all n large enough,

$$P\{\sup_x |f_n(x) - E\{f_n(x)\}| \geq \epsilon\}$$

$$\leq e^{-C_2 n h_n^m} + (1 + C_3/h_n^{1+m+m/\gamma})^m 2\, e^{-C_1 n h_n^m}$$

where

$$C_3 = (16C/\epsilon)(16K_1K_4/\epsilon)^{1/\gamma}.$$

Obviously, (4.30) implies that $\sum_{n=1}^{\infty} P\{\sup_x |f_n(x) - E\{f_n(x)\}| \geq \epsilon\} < \infty$ for all $\epsilon > 0$ so that, by the Borel-Cantelli lemma, $\sup_x |f_n(x) - E\{f_n(x)\}| \to 0$ wpl. Finally, (4.28) follows from (4.31) and theorem 4.7.

Q.E.D.

Proof of theorem 4.9

To prove theorem 4.9, we need the following auxiliary result. Let K be a Borel measurable function on $\mathbb{R}^m$ satisfying

(i) $0 \le K(x) \le M$, all $x \in \mathbb{R}^m$.

(ii) $K(x) \to 0$ as $\|x\| \to \infty$.

(iii) If $D = \{x : x \in \mathbb{R}^m, K \text{ is not continuous at } x\}$, then the closure of D (denoted by $\bar{D}$) has Lebesgue measure zero.

If a rectangle is a product of m finite intervals, then the following is true. For all $\epsilon > 0$ and $\delta > 0$ there exist integers N_1, N_2 and rectangles $A_1, \ldots, A_{N_1}, B_1, \ldots, B_{N_2}$ and positive numbers $\alpha_1, \ldots, \alpha_{N_1}, \rho$ such that with

$$K^*(x) = \sum_{i=1}^{N_1} \alpha_i I_{\{x \in A_i\}} \quad , x \in \mathbb{R}^m,$$

we have,

(iv) $|K^*(x) - K(x)| < \epsilon$ except on a set S.

(v) $S \subseteq \bigcup_{i=1}^{N_2} B_i$ where $\bigcup_{i=1}^{N_2} B_i$ has Lebesgue measure smaller than δ.

(vi) $0 \le K^*(x) \le M$ for all x.

(vii) $K^*(x) = 0$ outside a compact set $[-\rho, +\rho]^m$.

To see this, let $A_\rho = [-\rho, +\rho]^m$. By (ii), choose ρ such that $K(x) < \epsilon$ on $(A_\rho)^c$ where $(.)^c$ denotes the complement of a set. Let ν be the Lebesgue measure on $\mathbb{R}^m$. Note that $\nu(\bar{D} \cap A_\rho) \le \nu(\bar{D}) = 0$. Because $\bar{D} \cap A_\rho$ is measurable, there exists an open set O with $\bar{D} \cap A_\rho \subseteq O$ and $\nu(O) < \delta$. Since $\mathbb{R}^m$ is a separable metric space, we can find open rectangles $R_1, R_2, \ldots$ with $O = \bigcup_{i=1}^{\infty} R_i$. Now, $\{R_i\}$ is an open covering for the compact set $\bar{D} \cap A_\rho$ so that, by the Heine-Borel property, we can

find a finite subset $B_1,\ldots,B_{N_2}$ of $\{R_i, i=1,2,\ldots\}$ satisfying $\bar{D}\cap A_\rho \subseteq \bigcup_{i=1}^{N_2} B_i \triangleq E$. Clearly, $A_\rho \cap E^c$ is a compact set and since K is continuous on D^c, K must be uniformly continuous on $A_\rho \cap E^c$. Since $A_\rho \cap E^c$ is a finite number of rectangles, it is possible to partition $A_\rho \cap E^c$ into a disjoint union of N_1 rectangles each with diameter smaller than some $\theta>0$ where θ is picked so small that in each rectangle, $\sup K(x) - \inf K(x) < \epsilon$. Thus, $A_\rho \cap E^c = \bigcup_{i=1}^{N_1} A_i$. Note that one can always choose $N_1 \le (2N_2+2)^m + (1+ 2\rho/\theta)^m$. Pick one x_i in every A_i and let $\alpha_i = K(x_i)$. Clearly, $0 \le \alpha_i \le M$, $1 \le i \le N_1$. Let the α_i and A_i define K^*. By the disjointness of the A_i, $0 \le K^*(x) \le M$. Further $K^*=0$ on A_ρ^c and $|K(x)-K^*(x)| < \epsilon$ except possibly on a set S. Since $K(x) < \epsilon$ on A_ρ^c, it is clear that $S \subseteq A_\rho$. By construction we know that $S \subseteq (A_\rho \cap E^c)^c$. Therefore $S \subset E = \bigcup_{i=1}^{N_2} B_i$ and $\nu(E) < \delta$, proving (iv-vii).

By (4.31) and theorem 4.7 we need only show that $\sup_x |f_n(x)-E\{f_n(x)\}| \xrightarrow{n} 0$ wpl. We will show that there exist $C_1>0$, $C_2>0, C_3>0$ such that for all n large enough,

$$P\{\sup_x |f_n(x)-E\{f_n(x)\}| \ge \epsilon\} \le C_1 n^{C_2} e^{-C_3 n h_n^m}.$$

Of course, theorem 4.9 follows then immediately by (4.30) and the Borel-Cantelli lemma.

Let us first define $A_{(x,a)}$ where A is a Borel set from $\mathbb{R}^m$, $x \in \mathbb{R}^m$ and $a>0$:

$$A_{(x,a)} = \{y: y \in \mathbb{R}^m, y=x+at \text{ for some } t \in A\}.$$

Next, let $K_1=\sup_x K(x)$, $K_2=\sup_x \|x\|^m K(x)$ and $K_3=\sup_x f(x)$. Choose $\delta>0$ so small that $\delta < \epsilon/12 K_1 K_3$ and choose $\theta>0$ so small that $\theta < \epsilon/8 K_3 (2\rho)^m$ where ρ is defined in (4.33)(iii).

Because K satisfies the conditions (i-iii), we know that given $\delta>0$ and $\theta>0$ there exist integers N_1, N_2 and rectangles $A_1,\ldots,A_{N_1}$, $B_1,\ldots,B_{N_2}$ and numbers $a_1,\ldots,a_{N_1}$ such that the estimate

$$K^*(x) = \sum_{i=1}^{N_1} a_i I_{\{x\in A_i\}}$$

satisfies

1) $|K^*(x)-K(x)| \le \theta$ except on a set S.

2) $S\subset \bigcup_{i=1}^{N_2} B_i$, which has Lebesgue measure smaller than δ and is contained in $[-\rho,+\rho]^m$.

3) $K^*=0$ on $([-\rho,+\rho]^m)^c$.

4) $0\le K^*(x)\le K_1$ for all $x\in\mathbb{R}^m$.

Let $T=\bigcup_{i=1}^{N_2} B_i$. Then we have,

$$\begin{aligned}
&\sup_x |f_n(x)-E\{f_n(x)\}| \\
&\quad = \sup_x \left| \int h_n^{-m} K((y-x)/h_n)\, dF_n(y) - \int h_n^{-m} K((y-x)/h_n)\, dF(y) \right| \\
&\quad \le \sum_{i=1}^{3} \sup_x U_i(x)
\end{aligned}$$

where

$$U_1(x) = \left| \int h_n^{-m} K^*((y-x)/h_n)\, dF_n(y) - \int h_n^{-m} K^*((y-x)/h_n)\, dF(y) \right| ,$$

$$U_2(x) = \int h_n^{-m} |K((y-x)/h_n)-K^*((y-x)/h_n)|\, dF_n(y)$$

and

$$U_3(x) = \int h_n^{-m} |K((y-x)/h_n)-K^*((y-x)/h_n)|\, dF(y) .$$

Let $A^*=[-1,+1]^m$, $B_1^*=A^{*c}_{(x,\rho h_n)}$, $B_2^*=A^*_{(x,\rho h_n)}\cap S^c_{(x,h_n)}$ and $B_3^*=S_{(x,h_n)}$.

Now,

$$\sup_x U_3(x) \le \sum_{i=1}^{3} \sup_x \int_{B_i^*} h_n^{-m} |K((y-x)/h_n)-K^*((y-x)/h_n)|\, dF(y) .$$

Let ν denote the Lebesgue measure on $\mathbb{R}^m$. Then we have

$$\sup_x U_3(x) \le 0 + \theta K_3 h_n^{-m} \nu(A^*_{(x,\rho h_n)}) + K_1 K_3 h_n^{-m} \nu(S_{(x,h_n)})$$

$$\le \theta K_3 h_n^{-m} (2\rho h_n)^m + K_1 K_3 \delta h_n^m h_n^{-m} < \epsilon/8 + \epsilon/12 .$$

Let G_n be the class of all rectangles from $\mathbb{R}^m$ with maximal distance between any two points in any rectangle in G_n being $m 2\rho h_n$. Note that all $A_{i(x,h_n)}$, $B_{i(x,h_n)}$ and $A_{(x,\rho h_n)}$ belong to G_n.

Next,

$$\sup_x U_1(x) = \sup_x \left| \sum_{i=1}^{N_1} a_i h_n^{-m} (\mu_n(A_{i(x,h_n)}) - \mu(A_{i(x,h_n)})) \right|$$

$$\le \sum_{i=1}^{N_1} K_1 h_n^{-m} \sup_x |\mu_n(A_{i(x,h_n)}) - \mu(A_{i(x,h_n)})|$$

$$\le N_1 K_1 h_n^{-m} \sup_{A_0 \in G_n} |\mu_n(A_0) - \mu(A_0)|$$

and

$$\sup_x U_2(x) \le \sum_{i=1}^{3} \sup_x \int_{B_i^*} h_n^{-m} |K((y-x)/h_n) - K^*((y-x)/h_n)| \, dF_n(y)$$

$$\le 0 + \sup_x K_1 h_n^{-m} |\mu_n(T_{(x,h_n)}) - \mu(T_{(x,h_n)})|$$

$$+ \sup_x K_1 h_n^{-m} \mu(T_{(x,h_n)}) + \theta h_n^{-m} \sup_x \mu_{r_i}(A_{(x,\rho h_n)})$$

$$\le K_1 N_2 h_n^{-m} \sup_{A_0 \in G_n} |\mu_n(A_0) - \mu(A_0)| + K_1 h_n^{-m} K_2 \delta h_n^m$$

$$+ \theta h_n^{-m} \sup_x |\mu_n(A_{(x,\rho h_n)}) - \mu(A_{(x,\rho h_n)})| + \theta h_n^{-m} \sup_x \mu(A_{(x,\rho h_n)})$$

$$\le (K_1 N_2 + \theta) h_n^{-m} \sup_{A_0 \in G_n} |\mu_n(A_0) - \mu(A_0)| + \epsilon/12 + \theta K_2 (2\rho)^m$$

$$< (K_1 N_2 + \theta) h_n^{-m} \sup_{A_0 \in G_n} |\mu_n(A_0) - \mu(A_0)| + \epsilon/12 + \epsilon/8.$$

Collecting bounds we see that

$$\sup_x |f_n(x)-E\{f_n(x)\}|$$
$$< \epsilon/2 + (K_1N_1+K_1N_2+\theta)h_n^{-m} \sup_{A_0\in\mathcal{C}_n} |\mu_n(A_0)-\mu(A_0)|$$

and

$$P\{\sup_x |f_n(x)-E\{f_n(x)\}| \geq \epsilon\}$$
$$\leq P\{\sup_{A_0\in\mathcal{C}_n} |\mu_n(A_0)-\mu(A_0)| \geq h_n^m C_4\}$$

where $C_4 = \epsilon/2(K_1N_1+K_1N_2+\theta)$. Note that

$$\sup_{A\in\mathcal{C}_n'} \mu(A) \leq K_2(m4\rho h_n)^m$$

where $\mathcal{C}_n'$ is the class of all rectangles from $\mathbb{R}^m$ with maximal distance between any two points in any rectangle in $\mathcal{C}_n'$ being $4\rho m h_n$. By (3.29),

$$P\{\sup_{A_0\in\mathcal{C}_n} |\mu_n(A_0)-\mu(A_0)| \geq C_4h_n^m\}$$
$$\leq 4(1+2n)^{2m} e^{-nh_n^m(C_4^2/(64C_5+4C_4))} + 4ne^{-nh_n^mC_5/10}$$

for all n with $nh_n^mC_5 \geq 1$, $C_5h_n^m \leq 1/2$ and $nh_n^mC_4/(8C_5) \geq 1$, where $C_5 = K_2(4\rho m)^m$. This concludes the proof of theorem 4.9.

Q.E.D.

Proof of theorem 4.10

The proof of theorem 4.9 can be repeated with a few changes. First we know that we can find a K* as in theorem 4.9 (satisfying the properties 1)-4)). Without loss of generality, we can choose a ρ large enough so that

$$\int_\rho^\infty z^{m-1}u(z)\,dz < (\epsilon/16K_2)/2^{2m-1}.$$

Of course, with $\delta < \epsilon/12K_1K_3$ and $\theta < \epsilon/(8K_3(2\rho)^m)$, we have upon inspection of the proof of theorem 4.9

$$\sup_x |f_n(x)-E\{f_n(x)\}| \le \sum_{i=1}^{3} \sup_i U_i(x)$$

$$\le \sup_x \left(\int_{A^c_{(x,\rho h_n)}} h_n^{-m} K((y-x)/h_n)\, dF_n(y) \right.$$

$$\left. + \int_{A^c_{(x,\rho h_n)}} h_n^{-m} K((y-x)/h_n)\, dF(y) \right) + \varepsilon/2$$

$$+ (K_1 N_1 + K_1 N_2 + \theta) h_n^{-m} \sup_{A_0 \in \mathcal{G}_n} |\mu_n(A_0)-\mu(A_0)|$$

where $N_1, N_2, \mathcal{G}_n$ are defined as in the proof of theorem 4.9. For the first term on the right hand side of the last inequality, we argue as follows. We can upper bound it by

$$2 \sup_x \int_{A^c_{(x,\rho h_n)}} h_n^{-m} u(\|(y-x)/h_n\|)\, dF(y)$$

$$+ \sup_x \left| \int_{A^c_{(x,\rho h_n)}} h_n^{-m} u(\|(y-x)/h_n\|)\, dF_n(y) \right.$$

$$\left. - \int_{A^c_{(x,\rho h_n)}} h_n^{-m} u(\|(y-x)/h_n\|)\, dF(y) \right|$$

the first term of which is not greater than

$$2K_2 \int_\rho^\infty 2^{2m-1} z^{m-1} u(z)\, dz < \varepsilon/8$$

by the choice of ρ. Next, let $u'(x) = u(\|x\|) I_{\{\|x\| \ge \rho\}}$, let $K_4 = \sup_x u'(x)$ and let N_4 be the least integer not smaller than $16K_4/\varepsilon h_n^m$. Let $\mathcal{G}_n''$ be the class of all rectangles of $\mathbb{R}^m$ with maximal distance between any two points in every rectangle not greater than $\mathrm{Max}(2m\rho h_n,\ 2h_n u^{-1}(K_4/N_4))$. Let

$$S_{j,N_4} = \{x : x \in \mathbb{R}^m ;\ K_4(j-1)/N_4 < u'(x) \le K_4 j/N_4\}$$

and

$$T_{j,N_4} = \{x: x \in \mathbb{R}^m ; K_4(j-1)/N_4 < u'(x)\} = \bigcup_{i=j}^{N_4} S_{i,N_4}, \quad 1 \le j \le N_4.$$

Approximate u' by

$$u''(x) = \sum_{j=2}^{N_4} (K_4(j-1)/N_4)\, I_{\{x \in S_{j,N_4}\}}$$

so that for all x and n,

$$|u'(x) - u''(x)| \le K_4/N_4 \le \varepsilon h_n^m/16 .$$

We will show that

$$\sup_x \left| \int_{A^c_{(x,\rho h_n)}} h_n^{-m} u(\|(y-x)/h_n\|)\, dF_n(y) - \int_{A^c_{(x,\rho h_n)}} h_n^{-m} u(\|(y-x)/h_n\|)\, dF(y) \right|$$

$$\le \varepsilon/8 + 2K_4 h_n^{-m} \sup_{A_0 \in \mathcal{C}''_n} |\mu_n(A_0) - \mu(A_0)| .$$

Notice first that for all $1 \le j \le N_4 - 1$, $S_{j,N_4} = T_{j,N_4} \cap T^c_{j+1,N_4}$ and that every $T_{j,N_4}(x, h_n)$ is the difference of two concentric nested rectangles of $\mathbb{R}^m$ (we use the continuity and monotonicity of u' on $\{\|x\| \ge \rho\}$ and the fact that $u' = 0$ on $\{\|x\| < \rho\}$). Let $T_{j,N_4}(x, h_n) = T^1_{j,N_4}(x,h_n) \cap (T^2_{j,N_4}(x,h_n))^c$ where

$$T^1_{j,N_4}(x,h_n) = \{y: y \in \mathbb{R}^m ; K_4(j-1)/N_4 < u(\|(y-x)/h_n\|)\}$$

and

$$T^2_{j,N_4}(x,h_n) = A_{(x,\rho h_n)} .$$

It is clear that both rectangles belong to $\mathcal{C}''_n$. Indeed, if $j \ge 2$, then

$$T^1_{j,N_4}(x,h_n) = \{y: y\in\mathbb{R}^m ; \|y-x\| < h_n\, u^{-1}(K_4/N_4)\} \in \mathcal{C}''_n .$$

Thus,

$$\sup_x \left| \int_{A^c_{(x,\rho h_n)}} h_n^{-m} u(\|y-x\|/h_n) dF_n(y) - \int_{A^c_{(x,\rho h_n)}} h_n^{-m} u(\|y-x\|/h_n) dF(y) \right|$$

$$= \sup_x \left| \int h_n^{-m} u'(\|y-x\|/h_n)\, dF_n(y) - \int h_n^{-m} u'(\|y-x\|/h_n) dF(y) \right|$$

$$\le \sup_x \left| \int h_n^{-m} u''((y-x)/h_n) dF_n(y) - \int h_n^{-m} u''((y-x)/h_n)\, dF(y) \right|$$

$$+ \sup_x \int h_n^{-m} \left| u''((y-x)/h_n) - u'(\|y-x\|/h_n) \right| dF_n(y)$$

$$+ \sup_x \int h_n^{-m} \left| u''((y-x)/h_n) - u'(\|y-x\|/h_n) \right| dF(y)$$

$$\le 2\epsilon/16 + \sup_x K_4 h_n^{-m} \left| \sum_{j=2}^{N_4} (j-1) N_4^{-1} (\mu_n(S_{j,N_4}(x,h_n)) - \mu(S_{j,N_4}(x,h_n))) \right|$$

$$= \epsilon/8 + K_4 h_n^{-m} \sup_x \left| \sum_{j=2}^{N_4} N_4^{-1} (\mu_n(T_{j,N_4}(x,h_n)) - \mu(T_{j,N_4}(x,h_n))) \right|$$

$$\le \epsilon/8 + K_4 h_n^{-m} \sup_{x\,;\, j\ge 2} \left| \mu_n(T_{j,N_4}(x,h_n)) - \mu(T_{j,N_4}(x,h_n)) \right|$$

$$\le \epsilon/8 + 2K_4 h_n^{-m} \sup_{A_0\in\mathcal{C}''_n} \left| \mu_n(A_0) - \mu(A_0) \right|$$

which was to be shown.

Recalling from the proof of theorem 4.9 that $\mathcal{C}_n \subseteq \mathcal{C}''_n$, we collect bounds and obtain

$$\sup_x \left| f_n(x) - E\{f_n(x)\} \right|$$

$$\le 3\epsilon/4 + (K_1N_1 + K_1N_2 + \theta + 2K_4) h_n^{-m} \sup_{A_0\in\mathcal{C}''_n} \left| \mu_n(A_0) - \mu(A_0) \right|$$

and

$$P\{\sup_x \left| f_n(x) - E\{f_n(x)\} \right| \ge \epsilon\}$$

$$\le P\{ \sup_{A_0\in\mathcal{C}''_n} \left| \mu_n(A_0) - \mu(A_0) \right| \ge K_5 h_n^m \}$$

where

$$K_5 = \epsilon/4(K_1N_1+K_1N_2+\theta+2K_4) .$$

Now, note that if $c_n=\text{Max}(2m\rho h_n ; 2h_n u^{-1}(K_4/N_4))$ and $b_n=(4u^{-1}(\epsilon h_n^m/32))^m$, then

$$\sup_x \mu(S(x,2c_n)) \le K_2(2c_n)^m \le K_2 h_n^m b_n$$

for all n large enough in view of the choice of N_4, the monotonicity of u and $h_n \to 0$. From (3.29),

$$P\{ \sup_{A_0 \in \mathcal{C}''_n} |\mu_n(A_0)-\mu(A_0)| \ge K_5 h_n^m \}$$
$$\le 4(1+2n)^{2m} e^{-nh_n^{2m}K_5^2/(64K_2h_n^m b_n + 4K_5h_n^m)} + 4n e^{-nh_n^m K_2 b_n}$$

for all n large enough so that

(i) $4u^{-1}(K_4/N_4) \ge 4m\rho$,

(ii) $K_2 h_n^m b_n \le 1/2$,

(iii) $K_2 nh_n^m b_n \ge 1$, and

(iv) $nh_n^{2m}K_5^2 \ge 8K_2h_n^m b_n$.

Theorem 4.10 now follows from (4.35) and the Borel-Cantelli lemma.

Q.E.D.

Proof of theorem 4.11

Let $x \in Q(\mu)$ and let f be continuous at x. Let $T_j=(2\|X_j^x-x\|)^m$, $1 \le j \le n$. Thus, T_j is the volume of the sphere centered at x with radius $\|X_j^x-x\|$. Then, with an arbitrary $\epsilon>0$,

$$P\{|f_n(x)-f(x)| > \epsilon\}$$
$$\le P\{\sum_{j=k_{1n}}^{k_{2n}} \alpha_{jn} |j/nT_j - f(x)| > \sum_{j=k_{1n}}^{k_{2n}} \alpha_{jn}\epsilon \}$$

$$\leq (k_{2n}-k_{1n}+1) \sup_{k_{1n}\leq j\leq k_{2n}} P\{|j/nT_j - f(x)| > \epsilon\}.$$

Let $k_{1n}\leq j\leq k_{2n}$. Then, with $f(x)>\epsilon$,

$$\{|j/nT_j - f(x)| > \epsilon\} = \{T_j < j/n(f(x)+\epsilon)\} \cup \{T_j > j/n(f(x)-\epsilon)\}$$

and with $f(x)\leq\epsilon$,

$$\{|j/nT_j - f(x)| > \epsilon\} = \{T_j < j/n(f(x)+\epsilon)\}.$$

Assume first that $f(x)>\epsilon$. Let $K_1=(f(x)-\epsilon/2)/(f(x)-\epsilon)$, $K_2=\epsilon/4(f(x)-\epsilon)$, $K_3=(f(x)+\epsilon/2)/(f(x)+\epsilon)$ and $K_4=\epsilon/2(f(x)+\epsilon)$. Let

$$Y_i = I_{\{(2\|X_i-x\|)^m \leq j/n(f(x)-\epsilon)\}}, \quad 1\leq i\leq n.$$

It is clear that $Y_1,\ldots,Y_n$ are iid random variables and that

$$P\{Y_1=1\} = E\{Y_1\} = P\{(2\|X_1-x\|)^m \leq j/n(f(x)-\epsilon)\}$$

$$\in \left((j/n)(f(x)-\epsilon/2)/(f(x)-\epsilon),\ (j/n)(f(x)+\epsilon/2)/(f(x)-\epsilon)\right)$$

for all n large enough (where the "large enough" does not depend upon j). Indeed, since $k_{2n}/n \to 0$ and $j\leq k_{2n}$, we let N be so large that

$$(k_{2n}/n(f(x)-\epsilon))^{1/m}/2$$

is for all $n\geq N$ smaller than some $\delta>0$, where δ is such that $\|x-y\|<\delta$ implies that $|f(x)-f(y)|<\epsilon/2$ by the continuity of f at x. Obviously, for all $n\geq N$, we also have,

$$P\{(2\|X_1-x\|)^m \leq j/n(f(x)+\epsilon)\}$$

$$\in \left((j/n)(f(x)-\epsilon/2)/(f(x)+\epsilon),\ (j/n)(f(x)+\epsilon/2)/(f(x)+\epsilon)\right).$$

So,

$$P\{T_j > j/n(f(x)-\epsilon)\} \leq P\{\sum_{i=1}^{n} Y_i < j\}$$

$$\leq P\{\sum_{i=1}^{n}(Y_i - E\{Y_i\}) < j - j(f(x)-\epsilon/2)/(f(x)-\epsilon)\}$$

$$\leq P\{\sum_{i=1}^{n}(Y_i - E\{Y_i\})/n < -(k_{1n}/n)(\epsilon/2(f(x)-\epsilon))\}$$

$$\leq P\{\sum_{i=1}^{n}(Y_i - E\{Y_i\})/n < -K_2 k_{1n}/n\}$$

$$\leq e^{-n(K_2 k_{1n}/n)^2/(2(K_1+4K_2)k_{2n}/n + K_2 k_{1n}/n)}$$

$$\leq e^{-(k_{1n}^2/k_{2n})(K_2^2/(2K_1+9K_2))}$$

where we use the one-sided version of Bennett's inequality (Bennett, 1962) which is applicable because the Y_i are iid random variables with $E\{(Y_1 - E\{Y_1\})^2\} \leq E\{Y_1^2\} \leq (j/n)(f(x)+\epsilon/2)/(f(x)-\epsilon)$ $=(K_1+4K_2)j/n \leq (K_1+4K_2)k_{2n}/n$ and ess sup Y_1-ess inf $Y_1 \leq 1$. The inequalities are valid for $n \geq N$.

Similarly, with $Z_i = I_{\{(2\|X_i - x\|)^m \leq j/n(f(x)+\epsilon)\}}$, $1 \leq i \leq n$, we have that $E\{(Z_1 - E\{Z_1\})^2\} \leq E\{Z_1\} \leq K_3 k_{2n}/n$, ess sup Z_1-ess inf Z_1 ≤ 1 and $Z_1, \ldots, Z_n$ are independent identically distributed. Thus, by Bennett's inequality, for all $n \geq N$,

$$P\{T_j < j/n(f(x)+\epsilon)\} \leq P\{\sum_{i=1}^{n} Z_i \geq j\}$$

$$= P\{\sum_{i=1}^{n}(Z_i - E\{Z_i\}) \geq j - nE\{Z_1\}\}$$

$$\leq P\{\sum_{i=1}^{n}(Z_i - E\{Z_i\}) \geq j - j(f(x)+\epsilon/2)/(f(x)+\epsilon)\}$$

$$= P\{\sum_{i=1}^{n}(Z_i - E\{Z_i\}) \geq j\epsilon/2(f(x)+\epsilon)\}$$

$$\leq P\{\sum_{i=1}^{n}(Z_i - E\{Z_i\}) \geq K_4 k_{1n}/n\}$$

$$\leq e^{-n(K_4 k_{1n}/n)^2/(2K_3 k_{2n}/n + K_4 k_{1n}/n)}$$
$$\leq e^{-(k_{1n}^2/k_{2n})(K_4^2/(2K_3+K_4))} .$$

Thus, for all $n \geq N$,

$$\sup_{k_{1n} \leq j \leq k_{2n}} P\{ |j/nT_j - f(x)| \geq \epsilon \} \leq 2 e^{-C_1 k_{1n}^2/k_{2n}}$$

where $C_1 = \text{Min}\,(K_4^2/(2K_3+K_4)\ ,\ K_2^2/(2K_1+9K_2))$. So, for $n \geq N$ and $f(x) > \epsilon$,

$$P\{ |f_n(x) - f(x)| \geq \epsilon \} \leq 2(k_{2n} - k_{1n} + 1) e^{-C_1 k_{1n}^2/k_{2n}}$$
$$\leq 2n e^{-C_1 k_{1n}^2/k_{2n}}.$$

The theorem follows trivially. For the wp1 convergence part of the theorem, the Borel-Cantelli lemma is used. For $f(x) \leq \epsilon$ choose N^* so large that

$$(k_{2n}/n(f(x)+\epsilon))^{1/m}/2$$

is smaller than δ for all $n \geq N^*$. For such n, by a similar argument,

$$P\{ |f_n(x) - f(x)| \geq \epsilon \} \leq (k_{2n} - k_{1n} + 1) e^{-C_2 k_{1n}^2/k_{2n}}$$
$$\leq n e^{-C_2 k_{1n}^2/k_{2n}}$$

where $C_2 = K_4^2/(2K_3+K_4)$.

Q.E.D.

Proof of theorem 4.12

Let $\epsilon > 0$ be arbitrary and let $K_0 = \sup_x f(x)$, where f is a uniformly continuous density for μ. Pick $\delta > 0$ such that

$$\sup_x \sup_{y: \|y-x\| < \frac{1}{2}\delta^{1/m}} |f(y) - f(x)| < \epsilon/2$$

and find N so that for all $n \geq N$,

$$4k_{2n}/n\varepsilon < \delta$$

which is possible in view of $k_{2n}/n \overset{n}{\to} 0$. Now, with $T_j^x=(2\|X_j^x-x\|)^m$, $x\in\mathbb{R}^m$, $1\le j\le n$, we have

$$\begin{aligned}
&P\{\sup_x |f_n(x)-f(x)| > \varepsilon\} \\
&\le nP\{\bigcup_x \bigcup_{k_{1n}\le j\le k_{2n}} \{|j/nT_j^x - f(x)| > \varepsilon\}\} \\
&\le nP\{\bigcup_x \bigcup_{k_{1n}\le j\le k_{2n}} \{T_j^x < j/n(f(x)+\varepsilon)\}\} \\
&\quad + nP\{\bigcup_{x:f(x)>\varepsilon} \bigcup_{k_{1n}\le j\le k_{2n}} \{T_j^x > j/n(f(x)-\varepsilon)\}\} \\
&\le nP\{\bigcup_x \bigcup_{k_{1n}\le j\le k_{2n}} \{T_j^x < j/n(f(x)+\varepsilon\}\} \\
&\quad + nP\{\bigcup_{x:f(x)>\varepsilon} \bigcup_{k_{1n}\le j\le k_{2n}} \{T_j^x > j/n(f(x)-3\varepsilon/4)\}\}.
\end{aligned}$$

Let $S(x,r)=\{y: y\in\mathbb{R}^m, (2\|y-x\|)^m \le r\}$ where $x\in\mathbb{R}^m$ and $r\ge 0$. $S(x,r)$ is clearly a product of m intervals from $\mathbb{R}$. Let

$$\mathcal{G}_n = \{S(x,r) : x\in\mathbb{R}^m, r\le 4k_{2n}/n\varepsilon\}.$$

Note that r can be considered as the volume of the sphere $S(x,r)$. For all $n\ge N$, we have,

$$\begin{aligned}
&\bigcup_x \bigcup_{k_{1n}\le j\le k_{2n}} \{T_j^x < j/n(f(x)+\varepsilon)\} \\
&\subseteq \bigcup_x \bigcup_{k_{1n}\le j\le k_{2n}} \bigcup_{r\le k_{2n}/n\varepsilon} \{|\mu_n(S(x,r))-\mu(S(x,r))| > j\varepsilon/2n(f(x)+\varepsilon)\} \\
&\subseteq \bigcup_x \bigcup_{r\le k_{2n}/n\varepsilon} \{|\mu_n(S(x,r))-\mu(S(x,r))| > k_{1n}\varepsilon/2n(K_0+\varepsilon)\} \\
&\subseteq \bigcup_{A\in\mathcal{G}_n} \{|\mu_n(A)-\mu(A)| > k_{1n}\varepsilon/2n(K_0+\varepsilon)\}
\end{aligned}$$

where the only hard step is the second one.

Because $j/n(f(x)+\epsilon) < k_{2n}/n\epsilon < \delta$, we know that $T_j^x < j/n(f(x)+\epsilon)$ implies that there exists a sphere of $\mathbb{R}^m$ centered at x which contains at least j of the observations $X_1, \ldots, X_n$ while for all y in the same sphere, $|f(y)-f(x)| < \epsilon/2$. Further, the volume of the sphere cannot be greater than $k_{2n}/n\epsilon$. Call this sphere $S(x,\lambda)$ and note that $\mu_n(S(x,\lambda)) \geq j/n$ and $\mu(S(x,\lambda)) < (f(x)+\epsilon/2)j/n(f(x)+\epsilon)$ if j and x are fixed. This explains the second and crucial implication. In a similar fashion,

$$\bigcup_{x:f(x)>\epsilon} \bigcup_{k_{1n} \leq j \leq k_{2n}} \{T_j^x > j/n(f(x)-3\epsilon/4)\}$$

$$\subseteq \bigcup_{x:f(x)>\epsilon} \bigcup_{k_{1n} \leq j \leq k_{2n}} \bigcup_{r \leq 4k_{2n}/n\epsilon} \{|\mu_n(S(x,r)) - \mu(S(x,r))| > j\epsilon/4nK_0\}$$

$$\subseteq \bigcup_{x} \bigcup_{r \leq 4k_{2n}/n\epsilon} \{|\mu_n(S(x,r)) - \mu(S(x,r))| \geq k_{1n}\epsilon/4nK_0\}$$

$$\subseteq \bigcup_{A \in \mathcal{C}_n} \{|\mu_n(A) - \mu(A)| \geq k_{1n}\epsilon/4nK_0\}$$

where we used the fact that if for a fixed j and x, $T_j^x > j/n(f(x)-3\epsilon/4)$, then there exists a sphere $S(x,\lambda)$ with $\lambda < 4k_{2n}/n\epsilon < \delta$ and $\mu_n(S(x,\lambda)) < j/n$ and $\mu(S(x,\lambda)) \geq (f(x)-\epsilon/2)j/n(f(x)-3\epsilon/4) \geq (j/n)(1+\epsilon/4(f(x)-3\epsilon/4)) \geq (j/n)(1+\epsilon/4K_0)$. Combining bounds yields, for all $n \geq N$,

$$P\{\sup_x |f_n(x)-f(x)| > \epsilon\}$$

$$\leq n\Big(P\{\bigcup_{A \in \mathcal{C}_n} \{|\mu_n(A)-\mu(A)| > (k_{1n}/n)(\epsilon/2(K_0+\epsilon))\}\}$$

$$+ P\{\bigcup_{A \in \mathcal{C}_n} \{|\mu_n(A)-\mu(A)| > (k_{1n}/n)(\epsilon/4K_0)\}\}\Big)$$

$$\leq 2nP\{\bigcup_{A\in\mathcal{C}_n}\{|\mu_n(A)-\mu(A)|>(k_{1n}/n)(\varepsilon/4(K_0+\varepsilon))\}\}.$$

Let $K_2=\varepsilon/4(K_0+\varepsilon)$, $K_3=2^m K_0 4/\varepsilon$ and $K_4=K_2^2/(64K_3+4K_2)$. Then we can reason as follows. The maximum distance between any two points in any rectangle in $\mathcal{C}_n$ is $R_0=(4k_{2n}/n\varepsilon)^{1/m}$. Let $\mathcal{C}_n'$ be the class of all the rectangles from $\mathbb{R}^m$ for which the maximal distance between any two points in any rectangle is $2R_0$. Then

$$\sup_{A\in\mathcal{C}_n'}\mu(A)\leq K_0(2R_0)^m = K_0 2^m 4k_{2n}/n\varepsilon .$$

From (3.29), lemma 3.4 and lemma 3.6 we have

$$2nP\{\bigcup_{A\in\mathcal{C}_n}\{|\mu_n(A)-\mu(A)|\geq K_2k_{1n}/n\}\}$$
$$\leq 2n\,4(1+2n)^{2m}e^{-n(K_2k_{1n}/n)^2/(64K_3k_{2n}/n+4K_2k_{1n}/n)}$$
$$+8n^2e^{-nK_3k_{2n}/10n}$$

for all n large enough so that

(i) $nK_3k_{2n}/n\geq 1$

(ii) $n(K_2k_{1n}/n)^2\geq 8K_3k_{2n}/n$

and

(iii) $K_3k_{2n}/n\leq 1/2$.

We remark that it is possible to find such large n in view of $k_{2n}\overset{n}{\to}\infty$, $k_{2n}/n\overset{n}{\to}0$ and $k_{1n}^2/k_{2n}\overset{n}{\to}\infty$. Thus, for all n large enough,

$$P\{\sup_x|f_n(x)-f(x)|>\varepsilon\}$$
$$\leq 8n(1+2n)^{2m}e^{-K_4k_{1n}^2/k_{2n}}+8n^2e^{-K_3k_{2n}/10}.$$

Theorem 4.12 follows by the Borel-Cantelli lemma, $k_{2n}\geq k_{1n}^2/k_{2n}$ and (4.44). Q.E.D.

AD-A032 738 TEXAS UNIV AT AUSTIN DEPT OF ELECTRICAL ENGINEERING F/G 12/1
NONPARAMETRIC DISCRIMINATION AND DENSITY ESTIMATION.(U)
DEC 76 L P DEVROYE AF-AFOSR-2371-72

UNCLASSIFIED AFOSR-TR-76-1208 NL

2 OF 2
AD A032738

END

DATE FILMED
2-77

Chapter 5 NONPARAMETRIC DISCRIMINATION

5.1 Introduction

In chapter 2 we established a link between the convergence of δ_n to δ^* and of L_n to L^*. In particular, theorem 2.1 is of general interest to the statistician who wants to ascertain the asymptotic optimality of the discrimination rule. In this chapter we assume that the statistician has no a priori knowledge about the distribution of (X,θ), that is, he does not know that the distribution of (X,θ) belongs to any prespecified parametric family of distributions. The rules that are studied in this chapter are therefore referred to as nonparametric discrimination rules. For surveys on nonparametric discrimination, see, for instance, Cover and Wagner (1975). For discrimination in general, see Duda and Hart (1973), Fukunaga (1972), Nagy (1968), Ho and Agrawala (1968) and Sebestyen (1962).

Assume that all of the μ_i of (2.1) have densities f_i that are μ-almost everywhere continuous. In section 2.2 we showed how to construct an asymptotically optimal decision rule using consistent density estimates of the f_i. In sections 5.3 and 5.4 we discuss in more detail the asymptotic optimality of such two-step rules if the statistician chooses to use nonparametric density estimates such as the kernel estimate and the Loftsgaarden-Quesenberry estimate. A natural modification of these rules considerably simplifying their formulation is also introduced, and conditions for their asymptotic optimality are given.

It is quite possible that the statistician does not know that X has a density f and that he does not want to make any assumptions concerning the μ_i. In that case, he is not sure whether the two-step rules with kernel estimates are asymptotically optimal. However, even without restrictions on the μ_i it is possible to find an asymptotically

optimal discrimination rule, provided that there exist μ-almost everywhere continuous versions for the $P\{\theta=j \mid X\}$, $1 \le j \le M$. In the next section, we present a generalized nearest neighbor rule that is asymptotically optimal in such discrimination problems, which we will refer to as type C_3 discrimination problems.

5.2 A Generalized Nearest Neighbor Rule

One of the simplest and most thoroughly studied discrimination rules is the nearest neighbor rule (Fix and Hodges, 1951 ; Cover and Hart, 1967) which lets $\theta_{V_n, X} = \theta_i$ if X_i is the nearest neighbor to X among $X_1, \ldots, X_n$. It is well-known that the nearest neighbor rule is not asymptotically optimal (Cover and Hart, 1967; Wagner, 1971) but that a modified version, the k-nearest neighbor rule, can be asymptotically optimal if k is allowed to vary with n such that $k \overset{n}{\to} \infty$ and $k/n \overset{n}{\to} 0$ (Fix and Hodges, 1951). It is this result that we wish to generalize in this section.

Consider a probability distribution $v_n = (v_{n1}, \ldots, v_{nn})$, here called a weight vector . The rules we are interested in can be roughly described as follows. A weight v_{ni} is given to the state of the i-th closest neighbor to X. The estimate for the state of X is obtained by adding all the weights for each possible state value and selecting the one with the largest total weight. As we will see, however, some care must be taken in handling ties in distance.

Let $\|.\|$ be a norm on $\mathbb{R}^m$ and let the statistician attach numbers Z_i to the (X_i, θ_i) where the Z_i, $1 \le i \le n$, are random variables that are independent of (X, θ) and V_n. Let $V'_n = (X_1, \theta_1, Z_1), \ldots, (X_n, \theta_n, Z_n)$ be the enlarged data. The Z_i are used to break ties and we therefore require that all the Z_i take different values with probability one, e.g., $Z_i = i$, $1 \le i \le n$. Given $x \in \mathbb{R}^m$, the permutation V'^{x}_n of V'_n is thus with pro-

bability one well-defined. We will write

$$V_n'^{x}=(X_1^{x},\theta_1^{x},Z_1^{x}),\ldots,(X_n^{x},\theta_n^{x},Z_n^{x})$$

where $\|X_1^{x}-x\|\le\ldots\le\|X_n^{x}-x\|$ and $Z_i^{x}<Z_{i+1}^{x}$ whenever $\|X_i^{x}-x\| = \|X_{i+1}^{x}-x\|$, $1\le i\le n-1$. Let us include the Z_i in the definition of the conditional probability of error, say $L_n=P\{\theta_{V_n',X}\neq\theta \mid V_n'\}$ where $\theta_{V_n',X}$ is a random variable that is conditionally independent of θ given V_n' and X with

$$P\{\theta_{V_n',X}=j \mid V_n',X\} = \delta_{nj}(V_n',X)\ ,1\le j\le M,$$

and where $\delta_n=(\delta_{n1},\ldots,\delta_{nM})$ is our decision function, that is, a Borel measurable mapping from $(\mathbb{R}^m\times\{1,\ldots,M\}\times\mathbb{R})^n\times\mathbb{R}^m$ to $[0,1]^M$ such that $\sum_{j=1}^{M}\delta_{nj}=1$.

To each $(X_i^{x},\theta_i^{x},Z_i^{x})$ the statistician attaches a weight v_{ni}, $1\le i\le n$, and then computes the total weights that are given to each of the states j,

$$W_j^{x}=\sum_{i=1}^{n} v_{ni}\, I_{\{\theta_i^{x}=j\}}\ ,1\le j\le M\ . \tag{5.1}$$

His decision function δ_n then must satisfy

$$\delta_{nj}(V_n',x)=0 \text{ whenever } W_j^{x}<\max_{1\le i\le M} W_i^{x}\ ,1\le j\le M, x\in\mathbb{R}^m, \tag{5.2}$$

that is, he must always choose one of the state values with the largest weight. We prove the following theorem.

Theorem 5.1. If (2.1) defines a type C_3 discrimination problem and if $\{\delta_n\}$ is a decision rule satisfying (5.1)-(5.2) for all n, where

$$\sup_i v_{ni}\overset{n}{\to}0\ , \tag{5.3}$$

$$k_n/n \xrightarrow{n} 0 \tag{5.4}$$

and

$$\sum_{i=k_n+1}^{n} v_{ni} \xrightarrow{n} 0 \tag{5.5}$$

for some sequence of integers $\{k_n\}$, then $L_n \xrightarrow{n} L^*$ in probability. If, in addition,

$$\sum_{n=1}^{\infty} e^{-\alpha/\sup_i v_{ni}} < \infty \quad \text{for all } \alpha>0, \tag{5.6}$$

then $L_n \xrightarrow{n} L^*$ wpl. Remark that

$$(\log n) \sup_i v_{ni} \xrightarrow{n} \infty \tag{5.7}$$

is sufficient for (5.6).

The condition (5.3) insures that every weight v_{ni} is asymptotically negligible, while (5.5) makes the tail of the weight vector v_n asymptotically negligible. It seems natural to attach larger weights to nearer neighbors (i.e., $v_{n1} \geq v_{n2} \geq \ldots \geq v_{nn}$) but this is by no means necessary to insure the asymptotic optimality of the decision rule. Examples of sequences $\{v_n\}$ satisfying the conditions of theorem 5.1 are given below.

(i) rectangular weight vector. Let $1 \leq k_n \leq n$ for all n, and

$$v_{ni} = \begin{cases} 1/k_n & , 1 \leq i \leq k_n \\ 0 & , \text{otherwise} \end{cases}$$

where $k_n \xrightarrow{n} \infty$ and $k_n/n \xrightarrow{n} 0$. To satisfy (5.6), let additionally $\sum_{n=1}^{\infty} e^{-\alpha k_n} < \infty$ for all $\alpha>0$. Notice that with this choice of $\{v_n\}$, the decision rule reduces to the k_n-nearest neighbor rule (Cover and Hart, 1967).

(ii) triangular weight vector. Let $1 \leq k_n \leq n$ for all n, and let

$$v_{ni} = \begin{cases} 2(k_n - i + 1)/(k_n + k_n^2) & , 1 \le i \le k_n \\ 0 & , \text{otherwise} \end{cases}$$

where $k_n \overset{n}{\to} \infty$ and $k_n/n \overset{n}{\to} 0$. To satisfy (5.6), let additionally $\sum_{n=1}^{\infty} e^{-\alpha k_n} < \infty$ for all $\alpha > 0$.

(iii) exponential weight vector. Let

$$v_{ni} = (a_n/(1-(1+a_n)^{-n}))/(1+a_n)^i \quad , 1 \le i \le n,$$

where $a_n \in (0,\infty)$ for all n. The conditions (5.3)-(5.5) are satisfied if $a_n \overset{n}{\to} 0$ and $n a_n \overset{n}{\to} \infty$ (to see this, let $k_n \sim \sqrt{n/a_n}$ in (5.4) and (5.5)). Furthermore, $a_n \log n \overset{n}{\to} 0$ is sufficient for (5.6) and (5.7).

Let us briefly comment on the way distance ties are broken. For convenience, we introduced the random variables Z_i to uniquely define $V_n'^x$ given V_n'. However, theorem 5.1 remains valid if the decision rule is modified as follows. Let ξ_i^x , $1 \le i \le n$, be the number of X_ℓ's for which $\|X_\ell - x\| = \|X_i^x - x\|$. Thus, all the ξ_i^x are positive integer valued random variables. Let

$$W_j^x = \sum_{i=1}^{n} \sum_{k=1}^{n} v_{nk} I_{\{\theta_k^x = j\}} I_{\{\|X_k^x - x\| = \|X_i^x - x\|\}} / \xi_i^x , \qquad (5.8)$$

that is, if $\|X_1^x - x\| = \|X_2^x - x\| < \|X_3^x - x\|$ for instance, then both (X_1^x, θ_1^x) and (X_2^x, θ_2^x) carry an equal weight $(v_{n1} + v_{n2})/2$. Notice that the sums W_j^x in (5.8) are independent of the Z_i's. Thus it is possible to write $\theta_{V_n, X}$ as in chapter 2 and to speak of $L_n = P\{\theta_{V_n, X} \neq \theta \mid V_n\}$ and $\delta_n(V_n, x)$.

5.3 Distance Weighted Decision Rules

Assume that there exist μ-almost everywhere continuous densities $f_1, \ldots, f_M$ for the probability measures $\mu_1, \ldots, \mu_M$ in (2.1). In chapter 4 we have shown under what conditions the Parzen-

Rosenblatt density estimates f_{jn} converge to f_j, $1 \le j \le M$. Therefore, these density estimates can be used to construct an asymptotically optimal decision rule by following the technique given in section 2.2 (see (2.17)-(2.18)). As in (2.13), let N_{jn} be the number of observations X_i from V_n for which $\theta_i = j$. Let $\{h_n\}$ be a sequence from $(0,\infty)$ used in the Parzen-Rosenblatt estimator, then the two-step rule that uses the kernel estimator with kernel K is a sequence of decision functions δ_n satisfying

$$\delta_{nj}(V_n, x) = 0 \text{ whenever } W_j^x < \operatorname*{Max}_{1 \le i \le M} W_i^x, \quad 1 \le j \le M, x \in \mathbb{R}^m, \tag{5.9}$$

for all n where

$$W_j^x = \sum_{i=1}^{n} K((X_i - x)/h_{N_{jn}}) \, I_{\{\theta_i = j\}} / h_{N_{jn}}^m, \quad 1 \le j \le M.$$

Some of the in probability asymptotic properties of such decision rules can be found in Van Ryzin (1966). If the second derivatives of the f_j exist and are continuous and if the kernel K satisfies some additional regularity conditions, then Van Ryzin (1966) discusses the convergence to 0 of $P\{L_n - L^* > \epsilon_n\}$ for sequences $\{\epsilon_n\}$ with $\epsilon_n \overset{n}{\to} 0$. As an immediate corollary of theorems 2.2 and 4.2, we can state theorem 5.2, the wp1 version of which is new.

Theorem 5.2. If (2.1) defines a type C_1' discrimination problem, if K is a bounded probability density on $\mathbb{R}^m$ with $\|x\|^m K(x) \to 0$ as $\|x\| \to \infty$ and if $\{h_n\}$ is a sequence from $(0,\infty)$ with

$$h_n \overset{n}{\to} 0$$

and

$$n h_n^m \overset{n}{\to} \infty,$$

then any decision rule satisfying (5.9) for all n is asymptotically optimal. If in addition

$$\sum_{n=1}^{\infty} e^{-\alpha n h_n^m} < \infty \quad \text{for all } \alpha > 0$$

then $L_n - L^* \overset{n}{\to} 0$ wpl.

We note that (5.9) is not a very natural way of constructing a decision rule because the weight that is attached to each (X_i, θ_i) does not only depend upon $X_i - x$ but also on $N_{\theta_i n}$. This is undesirable because we want $\delta_n(V_n, x)$ to depend upon the (X_i, θ_i) for which $\|X_i - x\|$ is small. Consider the simpler decision rule defined as follows. Let δ_n be any decision function satisfying

$$\delta_{nj}(V_n, x) = 0 \text{ whenever } W_j^x < \operatorname{Max}_{1 \le i \le M} W_i^x, 1 \le j \le M, x \in \mathbb{R}^m, \quad (5.10)$$

for all n where

$$W_j^x = \sum_{i=1}^{n} K((X_i - x)/h_n)\, I_{\{\theta_i = j\}}, 1 \le j \le M,$$

and where K and $\{h_n\}$ are as in theorem 5.2. To prove the asymptotic optimality of such decision rules we can employ theorem 2.1. However, using an argument that differs only in the details from the proof of theorem 5.1, it is possible to obtain upper bounds for $P\{L_n - L^* > \epsilon\}$.

Theorem 5.3. If (2.1) defines a type C_1' discrimination problem, if K is a bounded, integrable mapping from $\mathbb{R}^m$ to $[0, \infty)$ with $\|x\|^m K(x) \to 0$ as $\|x\| \to \infty$ and if $\{h_n\}$ is a sequence from $(0, \infty)$ with $h_n \overset{n}{\to} 0$ and $n h_n^{m n} \to \infty$, then any decision rule satisfying (5.10) for all n is asymptotically optimal. If in addition

$$\sum_{n=1}^{\infty} e^{-\alpha n h_n^m} < \infty \quad \text{for all } \alpha > 0$$

then $L_n - L^* \overset{n}{\to} 0$ wpl. In particular, for every $\epsilon > 0$ there exist constants $K_1 > 0, N_1 > 0$ (both depending upon ϵ, (2.1) and K) such that for all $n \ge N_1$,

$$P\{L_n - L^* \geq \epsilon\} \leq (2/\epsilon) e^{-K_1 n h_n^m}. \tag{5.11}$$

We remark that the inequality (5.11) is only of academical importance because we can find distributions (2.1) defining type C_1' discrimination problems such that K_1 is arbitrarily small.

5.4 Two-Step Rules With Loftsgaarden-Quesenberry Density Estimates

Consider the two-step rule that is obtained by using the Loftsgaarden-Quesenberry density estimates $f_{jn}, 1 \leq j \leq M$, in (2.17)-(2.18). All we need to characterize this rule is a norm $\|.\|$ on $\mathbb{R}^m$ and a sequence $\{k_n\}$ of integers with $1 \leq k_n \leq n$. Let N_{jn} denote the number of observations X_i for which $\theta_i = j$. In particular, among those X_i, look for the c_{jn}-th nearest neighbor $X_{c_{jn}}^{x,j}$ to x where $x \in \mathbb{R}^m$ and $c_{jn} = k_{N_{jn}}$, and let U_{jn}^x be the Lebesgue measure of the sphere centered at x with $X_{c_{jn}}^{x,j}$ on its surface. Let $\{\delta_n\}$ be a decision rule satisfying

$$\delta_{nj}(V_n, x) = 0 \text{ whenever } W_j^x < \max_{1 \leq i \leq M} W_i^x, 1 \leq j \leq M, x \in \mathbb{R}^m, \tag{5.12}$$

for all n where

$$W_j^x = c_{jn} / U_{jn}^x \quad , 1 \leq j \leq M. \tag{5.13}$$

In particular, if $\|.\|$ is one of the L_r norms on $\mathbb{R}^m$ $(1 \leq r \leq \infty)$, then U_{jn}^x can be replaced by $\|X_{c_{jn}}^{x,j} - x\|^m, 1 \leq j \leq M$. From theorems 2.2 and 4.11 we can conclude, without proof, that the following is true.

Theorem 5.4. If (2.1) defines a type C_1' discrimination problem, if $\{\delta_n\}$ is a decision rule satisfying (5.12)-(5.13) for all n and if

$$k_n \xrightarrow{n} \infty \text{ and } k_n/n \xrightarrow{n} 0 \tag{5.14}$$

then $L_n - L^* \xrightarrow{n} 0$ in probability. If in addition

$$\sum_{n=1}^{\infty} e^{-\alpha k_n} < \infty \text{ for all } \alpha > 0 \tag{5.15}$$

then $L_n - L^* \xrightarrow{n} 0$ wpl.

Although this decision rule is computationally simple, it is unnatural because the $N_{jn}, 1 \le j \le M$, directly influence δ_n through the $c_{jn}, 1 \le j \le M$. This disrupts the local characteristics of the rule. The natural counterpart is obtained by replacing the c_{jn} by $k_n, 1 \le j \le M$. The resulting decision rule can be defined as follows. Let $\{\delta_n\}$ be a sequence of decision functions satisfying

$$\delta_{nj}(V_n, x) = 0 \text{ whenever } W_j^x > \min_{1 \le i \le M} W_i^x, 1 \le j \le M, x \in R^m, \tag{5.16}$$

for all n where

$$W_j^x = \begin{cases} \|X_{k_n}^{x,j} - x\| & \text{, if } N_{jn} \ge k_n \\ \infty & \text{, if } N_{jn} < k_n . \end{cases} \tag{5.17}$$

It is not hard to show that any decision rule satisfying (5.16)-(5.17) for all n is asymptotically optimal for all type C_1' discrimination problems. To prove this, observe that for every x, every event $\{\|X_i - x\| = \|X_\ell - x\|\}, i \ne \ell$, has zero probability. Thus, for M=2, we see that $L_n = L_n'$ wpl where L_n' is the conditional probability of error with the decision function discussed in section 5.2 with weight vector v_n where

$$v_{ni} = \begin{cases} 1/(2k_n - 1) & , 1 \le i \le 2k_n - 1 \\ 0 & , 2k_n \le i \le n , \end{cases}$$

and with any $Z_1, \ldots, Z_n$, provided that $2k_n - 1 \le n$. Applying theorem 5.1 thus yields,

Theorem 5.5. If (2.1) defines a type C_1' discrimination problem and if $\{k_n\}$ is a sequence of integers satisfying (5.14) with $1 \le k_n \le n$ for all n, then any decision rule satisfying (5.16)-(5.17) for all n is

asymptotically optimal. If in addition (5.15) holds, then $L_n \overset{n}{\to} L^*$ wpl as well.

We remark that both classes of decision rules of this section are not asymptotically optimal for all type C_3 discrimination problems. Let $m=1, M=2, \pi_1=2/3, \pi_2=1/3, k_n/n < 1/6$ and let μ_1 and μ_2 both put mass 1 at 0. Then, by Chebyshev's inequality, $P\{N_{1n} \geq k_n; N_{2n} \geq k_n\} \geq 1 - 2/4n(1/6)^2 = 1-18/n$. If $N_{1n} \geq k_n$ and $N_{2n} \geq k_n$, then $W_1^X = W_2^X = 0$ wpl. We pick δ_n such that $\delta_{n2}=1$ whenever $W_1^x = W_2^x$. Since $L^* = \frac{1}{3}$, we have that

$$L_n \geq I_{\{N_{1n} \geq k_n; N_{2n} \geq k_n\}} P\{\theta=1\}$$

and

$$P\{L_n - L^* \geq \tfrac{1}{3}\} \geq P\{N_{1n} \geq k_n; N_{2n} \geq k_n\} \geq 1 - 18/n \overset{n}{\to} 1.$$

A final remark is in order here. A comparison of the conditions of convergence in the theorems 5.3 and 5.5 shows that k_n plays the role of nh_n^m. Of course, this is not a total surprise. Indeed, for a given $x \in \mathbb{R}^m$ and kernel $K(y) = I_{\{\|y\| \leq 1\}}$, the decision functions satisfying (5.10) define a majority rule, that is, for each n, $\theta_{V_n, X}$ equals j if j is the state of the majority of the θ_i for which $\|X_i - X\| \leq h_n$. The number of observations for which $\|X_i - X\| \leq h_n$ is probabilistically proportional to nh_n^m. With the Loftsgaarden-Quesenberry version (5.16), the same majority rule is obtained except that the number of observations influencing $\theta_{V_n, X}$ is now probabilistically proportional to k_n.

5.5 Proofs

Assume throughout that the distribution (2.1) of (X, θ) is such that there exist μ-almost everywhere continuous versions $p_1, \ldots, p_M$ in (2.6). We first prove three lemmas that are of general interest in type C_3 discrimination problems. Consider first

$$g(x,\eta) = P\{\|X-x\| \leq \eta\}$$

where $\eta>0$, $\|.\|$ is a norm of $\mathbb{R}^m$, X is a random vector taking values in $\mathbb{R}^m$ with probability measure μ, and $x\in\mathbb{R}^m$. The following is true.

Lemma 5.1. For all $\eta>0$, $g(x,\eta)$ is a Borel measurable function of x, and

$$\lim_{b\to 0} P\{g(X,\eta)\leq b\}=0 .$$

Proof of lemma 5.1

It suffices to show that $P\{g(X,\eta)=0\}=0$. Let

$$S(x,t) = \{y : y\in\mathbb{R}^m, \|y-x\|\leq t\}$$

and note that $A= \{x : g(x,\eta)=0\} = \{x : P\{X\in S(x,\eta)\}=0\}$. Since $\mathbb{R}^m$ is separable, there exists a countable dense subset D of $\mathbb{R}^m$. Since D is dense, we can find for each x in A a $d(x)\in D$ such that $d(x) \in S(x,\eta/3)$. But

$$A \subset \bigcup_{\{y : y\in D, y=d(x) \text{ for some } x\in A\}} S(y,\eta/2)$$

which is a countable union of μ-null sets. This proves the second part of lemma 5.1. To prove that $g(x,\eta)$ is a Borel measurable function of x, we argue as follows. Let $\mathcal{B}^m$ be the σ-algebra of the Borel sets of $\mathbb{R}^m$ and let $\mathcal{B}^{2m}$ be the product σ-algebra of $\mathcal{B}^m$ and $\mathcal{B}^m$. Consider the product probability space $(\mathbb{R}^{2m}, \mathcal{B}^{2m}, \mu x \mu)$ and the following Borel set in $\mathcal{B}^{2m}$,

$$B = \{(x,y) : x,y\in\mathbb{R}^m, \|y-x\|\leq\eta\} .$$

We know that all the sections

$$B_z = \{(x,y) : x,y\in\mathbb{R}^m, x=z, \|y-x\|\leq\eta\} , z\in\mathbb{R}^m ,$$

are Borel sets of $\mathcal{B}^m$ (Loeve, 1963, pp.134) and that $\mu(B_z)$ is a Borel measurable function of z (Loeve, 1963, pp.135). Q.E.D.

Another quantity of some practical interest is

$$r(x,\eta) = \inf\Big(\|y-x\| : |p_i(y)-p_i(z)| \geq \eta \text{ for some } 1\leq i\leq M \text{ and some } y,z\in\mathbb{R}^m \text{ with } \|z-x\|\leq\|y-x\|\Big)$$

where $x\in\mathbb{R}^m$ and $\eta>0$. Note that r depends upon the particular version $(p_1,\ldots,p_M)$ that is chosen in (2.6). It is clear that for any $x,y\in\mathbb{R}^m$ and any $\eta>0$,

$$|r(x,\eta)-r(y,\eta)| \leq \|y-x\|$$

so that r is a Borel measurable function of x. Furthermore, the following is true in view of the Lebesgue dominated convergence theorem.

Lemma 5.2. For all $\eta>0$,

$$\lim_{b\to 0} P\{r(X,\eta)\leq b\} = 0$$

provided that the $p_1,\ldots,p_M$ in (2.6) that are used in the definition of r are μ-almost everywhere continuous.

Proof of lemma 5.2

Let C be the subset of $\mathbb{R}^m$ on which all the $p_i, 1\leq i\leq M$, are continuous. Since $P\{X\in C\} = 1$ by our choice of $p_1,\ldots,p_M$, we have by the Lebesgue dominated convergence theorem,

$$\lim_{b\to 0} \int_{\{x:\, x\in C, r(x,\eta)\leq b\}} \mu(dx) = \lim_{b\to 0} P\{X\in C; r(X,\eta)\leq b\}$$

$$= P\{X\in C; r(X,\eta)=0\} = P\{\emptyset\} = 0.$$

Q.E.D.

Another important tool is the inequality given in lemma 5.3. Given V_n and $x\in\mathbb{R}^m$, let $V_n^x=(X_1^x,\theta_1^x),\ldots,(X_n^x,\theta_n^x)$ be a permutation of V_n such that $\|X_1^x-x\|\leq\|X_2^x-x\|\leq\ldots\leq\|X_n^x-x\|$. Notice that V_n^x is not uniquely defined if ties occur. We assume that there is a method,

either deterministic or random, to break the ties and obtain V_n^x from V_n. For any tiebreaking method, possibly depending upon V_n, the following is true.

Lemma 5.3. If k_n is an integer with $1 \le k_n \le n$ and if $c>0, b>0$ and $g(x,c)>b$ where $x \in \mathbb{R}^m$, and if $k_n/n \le b/2$, then

$$P\{\|X_{k_n}^x - x\| > c\} \le e^{-nb/10}.$$

Proof of lemma 5.3

$$P\{\|X_{k_n}^x - x\| > c\} \le P\{\sum_{i=1}^{n} I_{\{\|X_i - x\| \le c\}} < k_n\}$$

$$\le P\{\frac{1}{n}\sum_{i=1}^{n} (I_{\{\|X_i - x\| \le c\}} - P\{\|X_i - x\| \le c\}) < k_n/n - b\}.$$

Since $k_n/n - b < -b/2$, we can upper bound the last term using Bennett's inequality (Bennett, 1962) by

$$e^{-n(b/2)^2/(2b+b/2)} = e^{-nb/10}.$$

Q.E.D.

Proof of theorem 5.1

For convenience, let M=2. Given a version (p_1, p_2) of (2.6) such that p_1 and p_2 are μ-almost everywhere continuous, we can define the following sets. Let

$$U = \{x : x \in \mathbb{R}^m ; p_1(x) + p_2(x) = 1\},$$

$$D^i = \{x : x \in \mathbb{R}^m ; p_i(x) = \text{Max}(p_1(x), p_2(x))\}, \quad i=1,2,$$

$$A(a) = \{x : x \in \mathbb{R}^m ; |p_1(x) - p_2(x)| \ge a\},$$

and

$$B(b,c,\eta) = \{x : x \in \mathbb{R}^m ; g(x,c)>b ; r(x,\eta)>c\}$$

where $a>0, b>0, c>0$ and $\eta>0$. Let $\epsilon>0$ be arbitrary. Then find an a small enough so that

$$P\{X \notin (D^1 \cap D^2) \cup A(a)\} < \epsilon/4$$

and find b and c small enough so that

$$P\{X \notin B(b,c,a/4)\} < \epsilon/4 .$$

Let $G = U \cap ((D^1 \cap D^2) \cup A(a)) \cap B(b,c,a/4)$ and note that $P\{X \notin G\} < \epsilon/2$. If $L_n(.)$ and $L^*(.)$ are defined as in (2.28),(2.30) by means of the same p_1, p_2, then we have

$$L_n(X) - L^*(X) \le I_{\{X \notin G\}} + (L_n(X) - L^*(X)) I_{\{X \in G\}} .$$

Taking conditional expectations yields, wp1,

$$L_n - L^* \le P\{X \notin G\} + E\{(L_n(X) - L^*(X)) I_{\{X \in G\}} \mid V_n\} .$$

By Markov's inequality and (2.28),(2.30),

$$\begin{aligned} P\{L_n - L^* \ge \epsilon\} &\le P\{E\{(L_n(X) - L^*(X)) I_{\{X \in G\}} \mid V_n\} \ge \epsilon/2\} \\ &\le (2/\epsilon)\, E\{(L_n(X) - L^*(X)) I_{\{X \in G\}}\} \\ &\le (2/\epsilon) \sup_{x \in G} E\{L_n(x) - L^*(x)\} \\ &\le (2/\epsilon) \sup_{x \in G} |p_1(x) - p_2(x)|\, E\{|\delta_1^*(x) - \delta_{n1}(V_n, x)|\} \\ &\le (2/\epsilon) \sup_{x \in U \cap A(a) \cap B(b,c,a/4)} E\{|\delta_1^*(x) - \delta_{n1}(V_n, x)|\} \\ &\le (2/\epsilon) \sup_{i=1,2} \sup_{x \in U \cap D^i \cap A(a) \cap B(b,c,a/4)} P\{\delta_{ni}(V_n, x) < 1\} . \end{aligned}$$

We will show that there exists an N such that for all $n \ge N$ and all x in $U \cap D^i \cap A(a) \cap B(b,c,a/4)$,

$$P\{\delta_{nj}(V_n, x) > 0\} \le e^{-nb/10} + e^{-(a^2/(128+8a))/\sup_i v_{ni}} , \qquad i=1,2 ;\ j=1,2 ;\ j \ne i .$$

This then would imply that for all n large enough,

$$P\{L_n - L^* \ge \epsilon\} \le \left(e^{-nb/10} + e^{-(a^2/(128+8a))/\sup_i v_{ni}}\right)(2/\epsilon)$$

from which theorem 5.1 follows by the arbitrariness of ϵ and the

Borel-Cantelli lemma.

So, we pick an x in $U \cap D^1 \cap A(a) \cap B(b,c,a/4)$ and we note that

$$P\{\delta_{n2}(V_n,x)>0\} \le P\{W_2^x \ge W_1^x\}$$

$$\le P\{\|X_{k_n}^x - x\| > c\} + P\{W_2^x \ge W_1^x ; \|X_{k_n}^x - x\| \le c\}$$

where $\{k_n\}$ is a sequence of integers satisfying (5.4)-(5.5). By lemma 5.3 and (5.4) we have $k_n/n < b/2$ for all n sufficiently large so that

$$P\{\|X_{k_n}^x - x\| > c\} \le e^{-n\,b/10}.$$

Next, introduce the random variables $Y_1^x, \ldots, Y_n^x$ where

$$Y_i^x = I_{\{\theta_i^x = 2\}} - I_{\{\theta_i^x = 1\}}, \quad 1 \le i \le n.$$

Note that on $\{\|X_{k_n}^x - x\| \le c\}$, we have

$$E\{Y_i^x \mid X_i^x\} = P\{\theta_i^x = 2 \mid X_i^x\} - P\{\theta_i^x = 1 \mid X_i^x\}$$

$$\le p_2(x) + a/4 - (p_1(x) - a/4) \le a/2 - a = -a/2 \quad \text{wp1}, 1 \le i \le k_n,$$

where we used the fact that $r(x,a/4) > c$ and $p_1(x) \ge p_2(x) + a$. So,

$$P\{W_2^x \ge W_1^x ; \|X_{k_n}^x - x\| \le c\} \le P\{\sum_{i=1}^{n} v_{ni} Y_i^x \ge 0 ; \|X_{k_n}^x - x\| \le c\}$$

$$\le P\{\sum_{i=1}^{n} v_{ni}(Y_i^x - E\{Y_i^x \mid X_i^x\}) > a/8 ; \|X_{k_n}^x - x\| \le c\}$$

$$+ P\{\sum_{i=1}^{k_n} v_{ni} E\{Y_i^x \mid X_i^x\} > -a/4 ; \|X_{k_n}^x - x\| \le c\}$$

$$+ P\{\sum_{i=k_n+1}^{n} v_{ni} E\{Y_i^x \mid X_i^x\} > a/8 ; \|X_{k_n}^x - x\| \le c\}.$$

Clearly, $|E\{Y_i^x \mid X_i^x\}| \le 1$ with probability one. Thus, by (5.5),

$$P\{\sum_{i=k_n+1}^{n} v_{ni} E\{Y_i^x \mid X_i^x\} > a/8\} = 0$$

for all n large enough. Also, since $E\{Y_i^x \mid X_i^x\} \le -a/2$ wp1 on $\{\|X_{k_n}^x - x\| \le c\}$ for all $1 \le i \le k_n$, we have

$$P\{\sum_{i=1}^{k_n} v_{ni} E\{Y_i^x \mid X_i^x\} > -a/4 \; ; \; \|X_{k_n}^x - x\| \le c\}$$

$$\le P\{-(a/2)\sum_{i=1}^{k_n} v_{ni} > -a/4\}$$

$$\le P\{(1-\sum_{i=k_n+1}^{n} v_{ni})a/2 < a/4\} = P\{\sum_{i=k_n+1}^{n} v_{ni} > \tfrac{1}{2}\} = 0$$

for all n large enough by (5.5). Finally, note that

$$P\{\sum_{i=1}^{n} v_{ni}(Y_i^x - E\{Y_i^x \mid X_i^x\}) > a/8\}$$

$$= E\{P\{\sum_{i=1}^{n} v_{ni}(Y_i^x - E\{Y_i^x \mid X_i^x\}) > a/8 \mid X_1^x, \ldots, X_n^x\}\}.$$

Notice that given $X_1^x, \ldots, X_n^x$, the $Y_i^x - E\{Y_i^x \mid X_i^x\}$ are independent, zero mean random variables. Notice further that given $X_1^x, \ldots, X_n^x$,

$$\sup_i \operatorname{ess\,sup} v_{ni}(Y_i^x - E\{Y_i^x \mid X_i^x\}) \le \sup_i v_{ni}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} E\{(v_{ni}(Y_i^x - E\{Y_i^x \mid X_i^x\}))^2 \mid X_1^x, \ldots, X_n^x\}$$

$$\le \frac{1}{n}\sum_{i=1}^{n} v_{ni}^2 \le \frac{1}{n} \sup_i v_{ni}$$

where the inequalities are with probability one. Therefore, by the one-sided inequality of Bennett for independent, but not identically distributed random variables (Bennett, 1962; Hoeffding, 1963; Fuk and Nagaev, 1971),

$$P\{\frac{1}{n}\sum_{i=1}^{n} v_{ni}(Y_i^x - E\{Y_i^x|X_i^x\}) > a/8n \mid X_1^x,\ldots,X_n^x\}$$
$$\leq e^{-n(a/8n)^2/((2/n)\sup_i v_{ni} + (a/8n)\sup_i v_{ni})}$$
$$= e^{-(a^2/(128+8a))/\sup_i v_{ni}}$$

with probability one. Taking expectations and collecting bounds yields the inequality given at the outset of the proof of theorem 5.1.

Q.E.D.

Proof of theorem 5.3

For convenience, let M=2. Find two μ-almost everywhere continuous densities f_1 and f_2 corresponding to μ_1 and μ_2. Let $f=\pi_1 f_1 + \pi_2 f_2$ and let for all $x\in R^m$ and $\eta>0$,

$$r(x,\eta) = \inf\{\|y-x\| : |\pi_i f_i(y)-\pi_i f_i(z)| \geq \eta \text{ for some } i=1,2 \text{ and some } y,z\in R^m \text{ with } \|z-x\|\leq\|y-x\|\}.$$

By lemma 5.2, since the $\pi_i f_i$ are μ-almost everywhere continuous, we know that

$$\lim_{b\to 0} P\{r(X,\eta) \leq b\} = 0.$$

Since r is Lipschitz, it is obviously a Borel measurable function of x. Of course, we also know that

$$\lim_{b\to\infty} P\{f(X) > b\} = 0.$$

Define the following sets,

$$D^i = \{x : x\in R^m, \pi_i f_i(x) = \text{Max}(\pi_i f_1(x), \pi_2 f_2(x))\}, i=1,2,$$

$$A(a) = \{x : x\in R^m, |\pi_1 f_1(x)-\pi_2 f_2(x)| \geq a\},$$

and

$$B(b,c,\eta) = \{x : x\in R^m, f(x)\leq b, r(x,\eta)>c\}$$

where $a>0, b>0, c>0$ and $\eta>0$.

Let $\epsilon>0$ be arbitrary. Find an $a>0$ small enough such that

$$P\{X \notin (D^1 \cap D^2) \cup A(a)\} < \epsilon/4$$

and find $b>0$ large enough and $c>0$ small enough such that

$$P\{X \notin B(b,c,a/4)\} < \epsilon/4.$$

Let $G=((D^1 \cap D^2) \cup A(a)) \cap B(b,c,a/4)$ and note that $P\{X \notin G\} < \epsilon/2$. If p_1 and p_2 are defined as in (2.15) and if $L_n(.)$ and $L^*(.)$ are defined by means of these p_1, p_2 (see (2.28), (2.30)), then we have

$$L_n(X)-L^*(X) \le I_{\{X \notin G\}} + (L_n(X)-L^*(X)) I_{\{X \in G\}}.$$

Taking conditional expectations yields, with probability one,

$$L_n - L^* \le P\{X \notin G\} + E\{(L_n(X)-L^*(X)) I_{\{X \in G\}} \mid V_n\}.$$

By Markov's inequality and (2.28), (2.30),

$$\begin{aligned}
P\{L_n - L^* \ge \epsilon\} &\le P\{E\{(L_n(X)-L^*(X)) I_{\{X \in G\}} \mid V_n\} \ge \epsilon/2\} \\
&\le (2/\epsilon) E\{(L_n(X)-L^*(X)) I_{\{X \in G\}}\} \\
&\le (2/\epsilon) \sup_{x \in G} E\{L_n(x)-L^*(x)\} \\
&\le (2/\epsilon) \sup_{x \in G} |p_1(x)-p_2(x)| E\{|\delta_1^*(x)-\delta_{n1}(V_n,x)|\} \\
&\le (2/\epsilon) \sup_{x \in A(a) \cap B(b,c,a/4)} E\{|\delta_1^*(x)-\delta_{n1}(V_n,x)|\} \\
&\le (2/\epsilon) \sup_{i=1,2} \sup_{x \in D^i \cap A(a) \cap B(b,c,a/4)} P\{\delta_{ni}(V_n,x) < 1\}.
\end{aligned}$$

We will show that there exists an $N_1>0$ such that for all $n \ge N_1$ and all x in $D^i \cap A(a) \cap B(b,c,a/4)$,

$$P\{\delta_{nj}(V_n,x)>0\} \le e^{-K_1 n h_n^m}, \quad i=1,2;\ j=1,2;\ j \neq i,$$

where $K_1>0$. This proves (5.11). Theorem 5.3 then follows from (5.11) and the Borel-Cantelli lemma.

Let $K_0 = \int K(x)\,dx$, $K_2 = \sup_x K(x)$ and $K_3 = \sup_x \|x\|^m K(x)$. Pick an x in $D^1 \cap A(a) \cap B(b,c,a/4)$ and define

$$W_i = K((X_i - x)/h_n)\,(I_{\{\theta_i=2\}} - I_{\{\theta_i=1\}}),\ 1 \le i \le n.$$

Then,

$$P\{\delta_{n2}(V_n,x) > 0\} \le P\{(\sum_{i=1}^{n} W_i)/n \ge 0\}$$

$$\le P\{\left(\sum_{i=1}^{n}(W_i - E\{W_i\})\right)/n \ge K_0 a h_n^m/8\} + P\{E\{W_1\} \ge -K_0 a h_n^m/8\}.$$

Further,

$$E\{W_1/h_n^m\} = \int h_n^{-m} K((y-x)/h_n)(\pi_2 f_2(y) - \pi_1 f_1(y))\,dy$$

$$= K_0(\pi_2 f_2(x) - \pi_1 f_1(x)) + \int h_n^{-m} K((y-x)/h_n)(\pi_2 f_2(y) - \pi_1 f_1(y))\,dy$$

$$- \int h_n^{-m} K((y-x)/h_n)(\pi_2 f_2(x) - \pi_1 f_1(x))\,dy$$

$$\le -K_0 a + \int_{\{y:\|y-x\| \le c\}} (a/4) h_n^{-m} K((y-x)/h_n)\,dy$$

$$+ \int_{\{y:\|y-x\| > c\}} h_n^{-m} K((y-x)/h_n)(\pi_2 f_2(y) - \pi_1 f_1(y) + b)\,dy$$

$$\le -K_0 a + K_0 a/4 + \int_{\{y:\|y-x\| > c\}} h_n^{-m} K((y-x)/h_n)\pi_2 f_2(y)\,dy$$

$$+ b \int_{\{y:\|y\| > c/h_n\}} K(y)\,dy .$$

Choose n so large that the last term is smaller than $K_0 a/4$ and that $h_n^{-m} K(y/h_n) \le K_0 a/4$ for all $\|y\| > c$. This is possible in view of the integrability of K, $\|y\|^m K(y) \to 0$ as $\|y\| \to \infty$ and $h_n \to 0$. We thus have that $E\{W_1/h_n^m\} \le -K_0 a/4$ for all n large enough uniformly over $D^1 \cap A(a) \cap B(b,c,a/4)$. Next, by Bennett's inequality (Bennett, 1962),

$$P\{\sum_{i=1}^{n}(W_i - E\{W_i\}) \ge n K_0 a h_n^m/8\}$$

$$\le \exp\left(-n(K_0 a h_n^m/8)^2 / (2E\{W_1^2\} + (\text{ess sup } W_1) K_0 a h_n^m/8)\right)$$

where we used the fact that the W_i are iid random variables. Note that

$$E\{K((X_1-x)/h_n)\} = \int K((y-x)/h_n)\, f(y)\, dy$$

$$\leq f(x)\, K_0 h_n^m + \int K((y-x)/h_n)\, (f(y)-f(x))\, dy$$

$$\leq bK_0 h_n^m + (a/4) K_0 h_n^m + \int_{\{y:\|y-x\|>c\}} K((y-x)/h_n)\, f(y)\, dy$$

$$\leq (b+a/4) K_0 h_n^m + K_3 h_n^m / c^m$$

$$\triangleq K_3' h_n^m .$$

Thus,

$$E\{W_1^2\} \leq K_2\, E\{K((X_1-x)/h_n)(I_{\{\theta_1=2\}} - I_{\{\theta_1=1\}})\} \leq K_2 K_3' h_n^m$$

and

$$\text{ess sup } W_1 \leq K_2.$$

Therefore, for all $n \geq N_1$ (N_1 independent of the x picked in $D^1 \cap A(a) \cap B(b,c,a/4)$),

$$P\{\delta_{n2}(V_n, x) > 0\} \leq e^{-K_1 n h_n^m}$$

where

$$K_1 = (K_0 a/8)^2 / (2K_2 K_3' + K_2 K_0 a/8)$$

and

$$K_3' = (b + a/4) K_0 + K_3/c^m.$$

Q.E.D.

Chapter 6 DISTRIBUTION-FREE ERROR ESTIMATION

6.1 Problem Formulation

After the statistician has picked a decision rule $\{\delta_n\}$ and after he has collected data V_n, he is of course interested in the performance of his rule, that is, he would like to estimate the conditional probability of error $L_n = P\{\theta_{V_n, X} \neq \theta \mid V_n\} = E\{1 - \delta_{n\theta}(V_n, X) \mid V_n\}$. Because the distribution of (X, θ) is unknown, there is no way to compute L_n. Instead, the statistician will have to use the data V_n to construct an estimate $\hat{L}_n$ of L_n. Ideally, he would like to obtain tight upper bounds for $P\{|L_n - \hat{L}_n| \geq \epsilon\}$ that do not depend upon the distribution of (X, θ). This would tell him how much confidence he can put in his estimate $\hat{L}_n$ without a priori knowledge of (2.1).

To illustrate how nontrivial this problem is consider the following example. Let $\delta_n = (\delta_{n1}, \ldots, \delta_{nM})$ be a constant. Then $L_n = 1 - \sum_{j=1}^{M} \pi_j \delta_{nj}$ can be estimated by replacing the π_j by their natural estimates N_j/n, $1 \leq j \leq M$, where the N_j are the number of X_i's for which $\theta_i = j$. So,

$$\hat{L}_n = 1 - \sum_{j=1}^{M} N_j \delta_{nj}/n$$

and

$$P\{|L_n - \hat{L}_n| \geq \epsilon\} \leq P\{\left|\sum_{j=1}^{M} (\pi_j - N_j/n)\, \delta_{nj}\right| \geq \epsilon\}$$

$$\leq \sum_{j=1}^{M} P\{|\pi_j - N_j/n| \geq \epsilon\} \leq 2M e^{-2n\epsilon^2}$$

by Hoeffding's inequality (Hoeffding, 1963). This bound is valid for all distributions of (X, θ) and all constant decision rules. Of course, constant decision rules are of no practical importance. A slightly better class of decision rules is the class for which

$$\delta_{nj}=0 \text{ whenever } N_j < \max_{1\le i\le M} N_i \ , 1\le j\le M.$$

Let for instance $M=2$ and $\delta_{n1}=a$ if $N_1=N_2$ where $0\le a\le 1$. Then, $L_n=\pi_1 I_{\{N_1<N_2\}}+\pi_2 I_{\{N_1>N_2\}}+(\pi_1(1-a)+\pi_2 a) I_{\{N_1=N_2\}}$. If $\hat{L}_n$ is the estimate obtained by replacing the unknown π_j by their estimates N_j/n, then

$$|L_n-\hat{L}_n| = 2\,|(\pi_1-N_1/n)(I_{\{N_1<N_2\}}-a\, I_{\{N_1=N_2\}})|$$

$$\le 2\,|\pi_1-N_1/n|$$

and

$$P\{|L_n-\hat{L}_n|\ge\epsilon\} \le 2e^{-n\epsilon^2/2}.$$

Thus, we are still able to find a distribution-free upper-bound for $P\{|L_n-\hat{L}_n|\ge\epsilon\}$. It is obvious that this class of decision rules can be improved upon. If $\{B_1, B_2, \dots\}$ is a partition of $\mathbb{R}^m$ and N_{jk} is the number of (X_i, θ_i)'s with $X_i\in B_k$ and $\theta_i=j$, we could let

$$\delta_{n1}(V_n, x) = \left.\begin{cases} 1 & \text{if } N_{1k}>N_{2k} \\ a_k & \text{if } N_{1k}=N_{2k} \\ 0 & \text{if } N_{1k}<N_{2k} \end{cases}\right\}, x\in B_k,$$

where $0\le a_k\le 1, k=1,2,\dots$. Unfortunately, it is impossible to find distribution-free upper-bounds for $P\{|L_n-\hat{L}_n|\ge\epsilon\}$ with this rule if $\hat{L}_n$ is obtained from L_n by replacing the unknown $P\{X\in B_k;\theta=j\}$ in the expression of L_n by their natural estimates N_{jk}/n. Assume that $M=2$, that $a_k=0$ for all k, that $\pi_2=0$ and that $P\{X\in B_k\}=1/2n$ for $1\le k\le 2n$ and $P\{X\in B_k\}=0$ otherwise. Then,

$$L_n = \sum_{k=1}^{\infty} P\{X\in B_k\}\, I_{\{N_{1k}=0\}} \ge n/2n = 1/2$$

and

$$\hat{L}_n = \sum_{k=1}^{\infty} (N_{1k}/n) I_{\{N_{1k}=0\}} = 0.$$

This means that for every n,

$$\sup_{(\pi_1,\ldots,\pi_M,\mu_1,\ldots,\mu_M)} P\{|L_n - \hat{L}_n| \geq 1/2\} = 1.$$

For every distribution of (X,θ), however, $\hat{L}_n$ converges to L_n wp1. Indeed, it is not hard to see that

$$P\{|L_n - \hat{L}_n| \geq \epsilon\} \leq P\{\sum_{k=1}^{\infty}\sum_{j=1}^{M} |P\{X \in B_k; \theta=j\} - N_{jk}/n| \geq \epsilon\}$$

which we know is upper-bounded by $C_1 e^{-C_2 n \epsilon^2}$ where $C_1 > 0$ and $C_2 > 0$ are constants depending upon ϵ and the distribution of (X,θ) (see proof of (3.31)). If the natural estimate $\hat{L}_n$ of L_n is not a distribution-free estimate (in the sense that for every n, there exists a $(\pi_1,\ldots,\pi_M,\mu_1,\ldots,\mu_M)$ such that $P\{|L_n - \hat{L}_n| \geq 1/2\} = 1$), then is it possible at all to find another distribution-free estimate for L_n ? The answer is yes. In sections 6.2 and 6.3 we will construct another estimate $\hat{L}_n$ of L_n for which $P\{|L_n - \hat{L}_n| \geq \epsilon\} \leq C/\sqrt{n}$ where C is a constant which depends upon M and ϵ only and which is <u>independent</u> of the distribution of (X,θ), of the a_k and of the partition $\{B_1, B_2, \ldots\}$.

In section 6.2 several estimates $\hat{L}_n$ are presented and some general purpose inequalities are proved relating $|L_n - \hat{L}_n|$ to δ_n. In the remaining sections, the error estimation problem is discussed separately for three large classes of decision rules,

(i) linear discrimination rules,

(ii) two-step rules

and

(iii) linear ordering rules and $\{k_n\}$-local rules.

All these rules except the $\{k_n\}$-local rules can be put into the

following format. Given the data and $x \in \mathbb{R}^m$, the statistician computes for each state j a real number $\psi_{nj}(V_n, x)$ that is his quantitative measure of the possibility that $\theta = j$ given V_n and given that $X = x$. Thereafter he compares all these numbers and lets

$$\delta_{nj} = 0 \text{ whenever } \psi_{nj} < \max_{1 \le i \le M} \psi_{ni}, \quad 1 \le j \le M. \tag{6.1}$$

Thus, the probability vector $(\delta_{n1}, \ldots, \delta_{nM})$ puts all its weight on those j for which $\psi_{nj} = \max_{1 \le i \le M} \psi_{ni}$. Consider the set of indices j for which $\psi_{nj} = \max_{1 \le i \le M} \psi_{ni}$, and let us say that this subset of $\{1, \ldots, M\}$ is the J-th of all possible 2^M subsets. We assume that the statistician has given himself 2^M probability vectors $\xi_n = (\xi_{n1}, \ldots, \xi_{nM})$ that are Borel measurable functions on $\mathbb{R}^m$, and that he picks the J-th probability vector for his decision function, say

$$\delta_n(V_n, x) = \xi_n(J, x) .$$

To insure that (6.1) is satisfied, we have to let $\xi_{nj} = 0$ whenever j does not belong to the J-th subset, that is, whenever $\psi_{nj} < \max_{1 \le i \le M} \psi_{ni}$.

For each class of rules we will find a nonparametric <u>distribution-free</u> estimate $\hat{L}_n$ of L_n, that is, an estimate with the property that for every $\epsilon > 0$, $P\{|L_n - \hat{L}_n| \ge \epsilon\}$ is upper-bounded by an appropriate function of n, ϵ, M, m and the parameters appearing in δ_n and $\hat{L}_n$ but which is independent of (2.1). Of course, 1 is always such a bound. We will say that an upper-bound for $P\{|L_n - \hat{L}_n| \ge \epsilon\}$ is <u>useful</u> for the given discrimination rule $\{\delta_n\}$ if the bound decreases to 0 uniformly over all distributions of (X, θ) as n tends to infinity.

If $\{\delta_n\}$ is an asymptotically optimal rule and $\hat{L}_n$ is a good estimate of L_n, then $\hat{L}_n$ can also be considered as an estimate of

L^* although it is obvious that $P\{|\hat{L}_n - L^*| \geq \varepsilon\}$ does not tend to 0 as $n \to \infty$ uniformly over all distributions (2.1) because $P\{|L_n - L^*| \geq \varepsilon\}$ does not converge to 0 uniformly over all distributions (2.1). We will therefore not study the properties of $\hat{L}_n$ as an estimate of L^*.

6.2 The Resubstitution, Deleted and Holdout Estimates

The oldest estimation technique discussed in the literature is the resubstitution estimate (see Toussaint (1974) for a survey of the literature on the estimation of the probability of error). The resubstitution estimate L_n^R is obtained by counting the number of errors if one estimates the states of all the observations X_i using δ_n and the data. Formally we have

$$L_n^R = \left(\sum_{i=1}^{n} I_{\{\theta_{V_n, X_i} \neq \theta_i\}} \right) / n \tag{6.2}$$

where the θ_{V_n, X_i}, $1 \leq i \leq n$, are, conditioned on V_n, independent random variables with

$$P\{\theta_{V_n, X_i} = j \mid V_n\} = \delta_{nj}(V_n, X_i), \quad 1 \leq j \leq M, 1 \leq i \leq n. \tag{6.3}$$

In general, L_n^R is an optimistic estimate of L_n because (X_i, θ_i) is contained in the data. By eliminating the randomization in (6.2)-(6.3) we obtain the strictly better estimate

$$L_n^{R'} = \left(\sum_{i=1}^{n} (1 - \delta_{n\theta_i}(V_n, X_i)) \right) / n . \tag{6.4}$$

Notice that $L_n^{R'} = E\{L_n^R \mid V_n\}$ and that $(E\{L_n^R \mid V_n\})^2 \leq E\{(L_n^R)^2 \mid V_n\}$ by Schwarz's inequality. Thus,

$$E\{(L_n - L_n^{R'})^2 \mid V_n\} = E\{L_n^2 \mid V_n\} - 2E\{L_n L_n^{R'} \mid V_n\} + E\{(L_n^{R'})^2 \mid V_n\}$$

$$\leq E\{(L_n - L_n^R)^2 \mid V_n\}. \tag{6.5}$$

It should be noted that the difference between both estimates is asymptotically negligible and has no bearing upon distribution-free results. To see this, notice that, with probability one,

$$E\{(L_n - L_n^R)^2 \mid V_n\} = E\{(L_n - L_n^{R'})^2 \mid V_n\} + E\{(L_n^R - L_n^{R'})^2 \mid V_n\}$$

$$\leq E\{(L_n - L_n^{R'})^2 \mid V_n\} + 1/4n \tag{6.6}$$

and

$$P\{|L_n^R - L_n^{R'}| \geq \epsilon \mid V_n\} \leq 2e^{-2n\epsilon^2}, \epsilon > 0 \tag{6.7}$$

by the conditional independence of the θ_{V_n, X_i} (given V_n), Chebyshev's inequality and Hoeffding's inequality (Hoeffding, 1963).

More recently the <u>deleted estimate</u> L_n^D has become very popular (Lachenbruch, 1967; Cover, 1969 ; Wagner, 1973 ; Rogers and Wagner, 1976). Let V_{ni} be the data with (X_i, θ_i) deleted, $1 \leq i \leq n$. Given δ_n, we construct a <u>deleted decision function</u> $\tilde{\delta}_n = (\tilde{\delta}_{n1}, \ldots, \tilde{\delta}_{nM})$ that is a Borel measurable mapping from $(\mathbb{R}^m \times \{1, \ldots, M\})^{n-1} \times \mathbb{R}^m$ to $[0,1]^M$ such that

$$\sum_{j=1}^{M} \tilde{\delta}_{nj} = 1 . \tag{6.8}$$

Conditioned on V_n, let θ_{V_{ni}, X_i}, $1 \leq i \leq n$, be independent random variables such that 1) given V_{ni} and X_i, θ_{V_{ni}, X_i} and θ_i are independent and 2)

$$P\{\theta_{V_{ni}, X_i} = j \mid V_n\} = \tilde{\delta}_{nj}(V_{ni}, X_i) , 1 \leq j \leq M, 1 \leq i \leq n. \tag{6.9}$$

The statistician can only gain by choosing $\tilde{\delta}_n$ close to δ_n although everything that will be said in this section carries through for any δ_n and any $\tilde{\delta}_n$. In the special sections on linear ordering rules and two-step rules, we will outline some very natural choices of $\tilde{\delta}_n$. We define L_n^D by

$$L_n^D = \left(\sum_{i=1}^{n} I_{\{\theta_{V_{ni}}, X_i \neq \theta_i\}} \right) / n . \tag{6.10}$$

A strictly better estimate is

$$L_n^{D'} = \left(\sum_{i=1}^{n} (1 - \tilde{\delta}_{n\theta_i}(V_{ni}, X_i)) \right) / n \tag{6.11}$$

where, arguing as with the resubstitution estimate, $L_n^{D'} = E\{L_n^D | V_n\}$,

$$E\{(L_n - L_n^{D'})^2 | V_n\} \leq E\{(L_n - L_n^D)^2 | V_n\} \leq E\{(L_n - L_n^{D'})^2 | V_n\} + 1/4n,$$

and

$$P\{|L_n^D - L_n^{D'}| \geq \epsilon | V_n\} \leq 2e^{-2n\epsilon^2}, \epsilon > 0. \tag{6.12}$$

The difference between both estimates is asymptotically negligible and has no bearing upon distribution-free results.

A decision function δ_n is <u>symmetric</u> if for every V_n and x, permutations of the data leave δ_n unchanged. A deleted decision function $\tilde{\delta}_n$ *is symmetric if for any* V_{n1} *and* x, $\tilde{\delta}_n$ *does not change* its value if the (X_i, θ_i) of V_{n1} are permuted. The reader will have no trouble to check that all the δ_n and $\tilde{\delta}_n$ that are discussed in this chapter are in fact symmetric. The following theorem, implicit in Rogers and Wagner (1976), is valid for such symmetric δ_n and $\tilde{\delta}_n$. It is the main tool in the development of distribution-free upper bounds for $P\{|L_n - L_n^D| \geq \epsilon\}$.

<u>Theorem 6.1</u>. If δ_n and $\tilde{\delta}_n$ are symmetric, then

$$E\{(L_n - L_n^D)^2\} \leq 1/n + 6E\{|\delta_{n\theta}(V_n, X) - \tilde{\delta}_{n\theta}(V_{n1}, X)|\} \tag{6.13}$$

where $(X, \theta), (X_1, \theta_1), \ldots, (X_n, \theta_n)$ are independent identically distributed random vectors. The inequality remains valid for $L_n^{D'}$.

As a corollary, we have for all $\epsilon > 0$,

$$P\{|L_n - L_n^D| \geq \epsilon\} \leq \left(1/n + 6E\{|\delta_{n\theta}(V_n, X) - \tilde{\delta}_{n\theta}(V_{n1}, X)|\}\right) / \epsilon^2 \tag{6.14}$$

and

$$E\{(L_n - L_n^D)^2\} \le 1/n + 6P\{\delta_{n\theta}(V_n, X) \ne \tilde{\delta}_{n\theta}(V_{n1}, X)\}. \tag{6.15}$$

Another historically important estimate is the holdout estimate L_n^H. Given V_n, let $T_n = (X_1, \theta_1), \ldots, (X_{n-s_n}, \theta_{n-s_n})$ and let $S_n = (X_{n-s_n+1}, \theta_{n-s_n+1}), \ldots, (X_n, \theta_n)$ where $1 \le s_n \le n-1$. T_n is called the training sequence and S_n is the testing sequence. Given δ_n, we construct a holdout decision function $\tilde{\delta}_n = (\tilde{\delta}_{n1}, \ldots, \tilde{\delta}_{nM})$ that is a Borel measurable mapping from $(\mathbb{R}^m \times \{1, \ldots, M\})^{n-s_n} \times \mathbb{R}^m$ to $[0,1]^M$ such that

$$\sum_{j=1}^{M} \tilde{\delta}_{nj} = 1. \tag{6.16}$$

Conditioned on T_n and X_{n-s_n+i}, $1 \le i \le n$, let the $\theta_{T_n, X_{n-s_n+i}}, \theta_{n-s_n+i}$, $1 \le i \le n$, be independent random variables taking values in $\{1, \ldots, M\}$ with

$$P\{\theta_{T_n, X_{n-s_n+i}} = j \mid V_n\} = \tilde{\delta}_{nj}(T_n, X_{n-s_n+i}),$$
$$1 \le j \le M \; ; \; 1 \le i \le s_n. \tag{6.17}$$

The statistician can only gain by choosing $\tilde{\delta}_n$ close to δ_n although everything that will be said in this section carries through for any δ_n and any $\tilde{\delta}_n$. The holdout estimate L_n^H is defined as follows.

$$L_n^H = \left(\sum_{i=1}^{s_n} I_{\{\theta_{T_n, X_{n-s_n+i}} \ne \theta_{n-s_n+i}\}} \right) / s_n \; . \tag{6.18}$$

The corresponding estimate without randomization is

$$L_n^{H'} = \sum_{i=1}^{s_n} (1 - \tilde{\delta}_{n\theta_{n-s_n+i}}(T_n, X_{n-s_n+i})) \; / s_n. \tag{6.19}$$

It is not hard to show that with probability one, the following inequalities hold true,

$$L_n^{H'} = E\{L_n^H | V_n\},$$

$$E\{(L_n - L_n^{H'})^2 | V_n\} \le E\{(L_n - L_n^H)^2 | V_n\}$$

$$\le E\{(L_n - L_n^{H'})^2 | V_n\} + 1/4s_n, \tag{6.20}$$

and

$$P\{|L_n^H - L_n^{H'}| \ge \epsilon \mid V_n\} \le 2e^{-2s_n\epsilon^2}, \epsilon > 0.$$

Note that if $s_n \overset{n}{\to} \infty$, the difference between both estimates becomes asymptotically negligible. The following theorem will be very useful in the development of distribution-free upper-bounds for $P\{|L_n - L_n^H| \ge \epsilon\}$.

Theorem 6.2. For all decision functions δ_n and holdout decision functions $\tilde{\delta}_n$ and for all $\epsilon > 0$,

$$E\{(L_n - L_n^H)^2\} \le 1/2 s_n + 2E\{|\delta_{n\theta}(V_n, X) - \tilde{\delta}_{n\theta}(T_n, X)|^2\}$$

$$\le 1/2 s_n + 2E\{|\delta_{n\theta}(V_n, X) - \tilde{\delta}_{n\theta}(T_n, X)|\}, \tag{6.21}$$

$$P\{|L_n - L_n^H| \ge \epsilon\} \le 2e^{-s_n\epsilon^2/2}$$

$$+ (4/\epsilon^2) E\{|\delta_{n\theta}(V_n, X) - \tilde{\delta}_{n\theta}(T_n, X)|^2\} \tag{6.22}$$

and

$$P\{|L_n - L_n^H| \ge \epsilon\} \le 2e^{-s_n\epsilon^2/2}$$

$$+ (2/\epsilon) E\{|\delta_{n\theta}(V_n, X) - \tilde{\delta}_{n\theta}(T_n, X)|\}. \tag{6.23}$$

Theorem 6.2 remains valid for $L_n^{H'}$.

Notice the similarity between theorem 6.1 for the deleted estimate and the inequality (6.21) for the holdout estimate. Theorems 6.1 and 6.2 imply that in the search for upper bounds for $E\{(L_n - L_n^D)^2\}$ and $E\{(L_n - L_n^H)^2\}$ it suffices to find upper bounds for $E\{|\delta_{n\theta}(V_n, X) - \tilde{\delta}_{n\theta}(V_{n1}, X)|\}$ and $E\{|\delta_{n\theta}(V_n, X) - \tilde{\delta}_{n\theta}(T_n, X)|\}$. This is

exactly what we will attempt to do in the following sections.

6.3 Two-Step Rules

In chapters 2 and 5 we defined two-step rules as rules that are constructed in two stages. In the first stage an estimate is constructed for some density or some atomic measure, and in the second stage these estimates are employed to find a decision function. In section 5 we have pointed out how the density estimate can be modified so that a more natural two-step rule is obtained. These natural two-step rules are the object of this section. We say that $\{\delta_n\}$ is a two-step rule if all the ψ_{nj} satisfy

$$\psi_{nj}(V_n, x) = \sum_{i=1}^{n} w_j^n(X_i, x)\, I_{\{\theta_i = j\}}, \quad 1 \le j \le M, \tag{6.24}$$

where the w_j^n are Borel measurable mappings from $\mathbb{R}^m \times \mathbb{R}^m$ to $\mathbb{R}$, $1 \le j \le M$, $n \ge 1$. If

$$w_j^n(y, x) = K((y-x)/h_n), \quad x, y \in \mathbb{R}^m, \quad 1 \le j \le M, \tag{6.25}$$

where $\{h_n\}$ is a sequence of positive numbers and K is a Borel measurable mapping from $\mathbb{R}^m$ to $[0, \infty)$, then we obtain the natural modification of the distance weighted decision rule of section 5.3. In particular, if $K(x) = I_{\{\|x\| \le 1\}}$, then we obtain a majority rule over all the (X_i, θ_i) for which $\|X_i - x\| \le h_n$ (Fix and Hodges, 1951; Sebestyen, 1962). The histogram decision rule of section 2.3 is also a two-step rule with the definition (6.24). Indeed, if $\{B_1, B_2, \ldots\}$ is a countable partition of $\mathbb{R}^m$, and if

$$w_j^n(y, x) = \sum_{k=1}^{\infty} I_{\{y \in B_k\}} I_{\{x \in B_k\}}, \quad x, y \in \mathbb{R}^m, \quad 1 \le j \le M, \tag{6.26}$$

the resulting decision function lets $\theta_{V_n, X} = j$ if X takes a value in B_k and if the majority of the observations X_i with $X_i \in B_k$ have

states $\theta_i = j$.

The property we will exploit in this section is that every ψ_{nj} is the sum of n independent identically distributed random variables. We will derive upper-bounds for $P\{|L_n^D - L_n| \geq \epsilon\}$ and $P\{|L_n^H - L_n| \geq \epsilon\}$ that depend upon n, ϵ, M and c_n where $c_n \geq 1$ is a characteristic of the functions $w_j^n, 1 \leq j \leq M$. We say that c_n is the ratio range of the w_j^n if the range of all the w_j^n is contained in $\{0\} \cup [a_n, b_n]$ for some $0 < a_n \leq b_n < \infty$ with $b_n/a_n = c_n$. The interesting feature of these bounds is that they do not depend upon the distribution of (X, θ) or upon the nature of the mappings $w_j^n, 1 \leq j \leq M$, provided that their ratio range is c_n or less. We will see that for every n we can find simple functions w_j^n with a finite ratio range c_n and a distribution of (X, θ) such that

$$P\{|L_n^D - L_n| \geq 1/3\} > 1/e^2.$$

So, the knowledge of c_n is sufficient for the statistician to obtain distribution-free upper-bounds for $P\{|L_n^D - L_n| \geq \epsilon\}$ and no distribution-free upper-bound exists that holds uniformly over all ratio ranges. The yet unsolved problem is that if the w_j^n are such that $c_n = \infty$, can one still find distribution-free upper-bounds for $P\{|L_n^D - L_n| \geq \epsilon\}$ that do depend upon some characteristic other than the ratio range?

The situation is even worse for the resubstitution estimate L_n^R. We will see that for the most common mappings $w_j^n, 1 \leq j \leq M$, with $c_n = 1$, it is possible to find a distribution of (X, θ) such that

$$P\{|L_n^R - L_n| \geq 1/4\} = 1.$$

Because the statistician does not know $\pi_1, \ldots, \pi_M, \mu_1, \ldots, \mu_M$, he will in those simple cases never be sure, no matter how large n is, whether the resubstitution estimate L_n^R is close to L_n. The conclusion

therefore is that for two-step rules, the deleted estimate seems at this moment to be the best choice as an estimate of L_n.

Let the w_j^n be as in (6.25) with $K(x)=I_{\{\|x\|\le 1\}}$, or as in (6.26) with any choice of $\{B_1, B_2, \dots\}$. In both cases the ratio range is $c_n=1$. However, for both trivial examples and for all the possible mappings ξ_n, it is possible to find distributions of (X,θ) such that $L_n - L_n^R \ge 1/4$.

Theorem 6.3. Let $w_j^n, 1\le j\le M$, be as in (6.25) with $K(x)=I_{\{\|x\|\le 1\}}$ or as in (6.26) with any choice of $\{B_1, B_2, \dots\}$, and let ξ_n be arbitrary. Then it is possible to find a distribution of (X,θ) such that $L_n - L_n^R \ge 1/4$.

Theorem 6.3 implies that even for the simplest two-step rule, i.e., one for which all the w_j^n take values in $\{0,1\}$, $1\le j\le M$, $n\ge 1$,

$$\sup_{(\pi_1,\dots,\pi_M,\mu_1,\dots,\mu_M)} P\{|L_n^R - L_n| \ge 1/4\} = 1, \quad n\ge 1. \tag{6.27}$$

Let us turn now to the deleted estimate. To define the deleted estimate, we first have to construct a deleted decision function $\tilde{\delta}_n=(\tilde{\delta}_{n1},\dots,\tilde{\delta}_{nM})$ that is close to δ_n. The statistician knows the w_j^n and ξ_n he is using. Thus, he can approximate ψ_n by $\psi_n'=(\psi_{n1}',\dots,\psi_{nM}')$ where

$$\psi_{nj}'(V_{ni},x) = \sum_{\substack{k=1\\k\ne i}}^{n} w_j^n(X_k,x)\, I_{\{\theta_k=j\}}, \quad 1\le j\le M, 1\le i\le n. \tag{6.28}$$

Let the J'-th subset of $\{1,\dots,M\}$ be the subset of indices j for which $\psi_{nj}' = \operatorname{Max}_{1\le i\le M} \psi_{ni}'$ and let $\tilde{\delta}_{ni}(V_{ni},x)=\xi_n(J',x)$. Because the same w_j^n and ξ_n are used in the construction of δ_n and $\tilde{\delta}_n$, we may expect that δ_n and $\tilde{\delta}_n$ are close to each other. The limited power

of the deleted estimate alluded to above is because of the following property.

Theorem 6.4. There exists a two-step rule $\{\delta_n\}$ with range $w_j^n \subset [0,1]$, $1 \le j \le M, n \ge 1$, such that for all n,

$$P\{|L_n^D - L_n| \ge 1/3\} > 1/e^2 \tag{6.29}$$

for some distribution of (X, θ).

The main result for the deleted estimate is the following.

Theorem 6.5. Let $\{\delta_n\}$ be any two-step rule where c_n is the ratio range of the $w_j^n, 1 \le j \le M$. Then, for all $\epsilon > 0$ and for all $\pi_1, \ldots, \pi_M$, $\mu_1, \ldots, \mu_M$,

$$P\{|L_n^D - L_n| \ge \epsilon\} \le e^{-2}\left(2/n + 6C_2(M-1)c_n/\sqrt{n}\right) \tag{6.30}$$

where

$$C_2 = 4/\sqrt{n} + 4(1 + 14/3\sqrt{2\pi} + 64C_1)\sqrt{2}$$

and where C_1 is the universal constant of the Berry-Esseen theorem (e.g., Zolotarev(1966) reports that $C_1 \le 1.322$).

In the proof of theorem 6.5, we use a uniform version of the Berry-Esseen central limit theorem to show that for all distributions of (X, θ),

$$E\{(L_n^D - L_n)^2\} \le 2/n + 6C_2(M-1)c_n/\sqrt{n}\,. \tag{6.31}$$

We do not pretend that the constant C_2 in (6.30), (6.31) is the smallest possible and indeed, it is possible, by further restricting the class of rules $\{\delta_n\}$, to obtain much smaller constants. However, it turns out that the bounds in (6.30), (6.31) are the best possible with regard to their dependence on n. In particular, it is possible to show the following.

Theorem 6.6. Let $M=2$, and $0<\varepsilon<1/2$. There exists a two-step rule with $c_n=1$ $(n\geq 1)$ such that for all even n,

$$\sup_{(\pi_1,\ldots,\pi_M,\mu_1,\ldots,\mu_M)} P\{|L_n^D-L_n|\geq\varepsilon\}\geq 1/\sqrt{2n} \qquad (6.32)$$

and

$$\sup_{(\pi_1,\ldots,\pi_M,\mu_1,\ldots,\mu_M)} E\{(L_n^D-L_n)^2\}\geq 1/4\sqrt{2n}\ . \qquad (6.33)$$

Let us now briefly discuss the use of L_n^D for the distance weighted decision rules of section 5.3. Consider first kernels K taking values 0 and 1. Then, no matter how the statistician chooses K, the bounds (6.30) and (6.31) are valid with $c_n=1$, independent of the choice of $\{h_n\}$. If K takes integer values, say $0,1,2,\ldots,N$, then, independent of h_n and the actual form of K, the bounds (6.30) and (6.31) are applicable with $c_n=N$. Unfortunately, some interesting kernels such as the gaussian have a compact range of the form $[0,\beta]$ and by theorem 6.4, we may suspect that L_n^D is not a good distribution-free estimate of L_n for such two-step rules. One way out for the statistician is to slightly change the decision rule. If range $K\subseteq[0,1]$, then he could replace K by $K'=K+d_n$ in (5.21) where $d_n>0$. In that case, the bounds (6.30) and (6.31) are valid with $c_n=(1+d_n)/d_n$. This modified rule is asymptotically equivalent in form to the original rule if $d_n\overset{n}{\to}0$, and the bound (6.31) is useful if $nd_n^2\overset{n}{\to}\infty$. It remains to be shown that the modified rule is asymptotically optimal under some condition on $\{d_n\}$ that does not contradict the condition $nd_n^2\overset{n}{\to}\infty$. Another modification would be to replace K in (5.21) by $K'=K\,I_{\{K\geq d_n\}}$ in which case $c_n=1/d_n$ if $d_n\leq 1$.

To illustrate the fact that other bounds can be found if the statistician has some a priori knowledge about the distribution of

(X,θ), consider the following example. Let $\mu_1,\ldots,\mu_M$ all put mass 0 outside $[-r,+r]^m$ where r is known to the statistician. Let K be a given kernel with a range of values on [0,1] and let the w_j^n be as in (6.25). It is easy to see that (6.30), (6.31) can be applied in this case with

$$c_n = 1/\inf_{x\in[-2r,+2r]^m} K(x/h_n).$$

Consider the case that K is bell-shaped, i.e. K is a monotonically nonincreasing function of $\|x\|$, say $K(x)=1/(\beta+\gamma\|x\|^{m+\alpha})$, where $\alpha>0, \beta\geq 1, \gamma>0$. In that case, (6.30) and (6.31) are applicable with

$$c_n = \beta+\gamma(2rm/h_n)^{m+\alpha}.$$

The holdout estimate L_n^H can also be used as a distribution-free estimate of L_n. We will see that the bounds for $P\{|L_n^H-L_n|\geq\epsilon\}$ are larger than the ones obtained for the deleted estimate. On the other hand, some statisticians may be attracted to the holdout estimate because of its simplicity. Most of the discussion for L_n^D can be repeated for L_n^H. We will only state how to choose the holdout decision function $\tilde{\delta}_n$ and then obtain a distribution-free upper bound for $P\{|L_n^H-L_n|\geq\epsilon\}$.

Given T_n and $x\in\mathbb{R}^m$, let $\tilde{\delta}_n(T_n,x)=\delta_n(J',x)$ where the J'-th subset of $\{1,\ldots,M\}$ is such that for every index j in the subset, $\psi'_{nj} < \max_{1\leq i\leq M} \psi'_{ni}$ where $\psi'_n=(\psi'_{n1},\ldots,\psi'_{nM})$ is defined as ψ_{n-s_n} but with the mappings $w_j^n, 1\leq j\leq M$. Thus,

$$\psi'_{nj}(T_n,x) = \sum_{i=1}^{n-s_n} w_j^n(X_i,x)\, I_{\{\theta_i=j\}}, \quad 1\leq j\leq M. \tag{6.34}$$

The main result for the holdout estimate is the following.

Theorem 6.7. Let $\{\delta_n\}$ be any two-step rule where c_n is the ratio range of the $w_j^n, 1 \le j \le M$. Then for all n, all $\epsilon > 0$ and all distributions of (X, θ),

$$P\{|L_n^H - L_n| \ge \epsilon\} \le 2e^{-s_n \epsilon^2/2} + 2C_2(M-1)c_n s_n/\sqrt{n-s_n+1} \tag{6.35}$$

and

$$E\{(L_n^H - L_n)^2\} \le 1/2s_n + 2C_2(M-1)c_n s_n/\sqrt{n-s_n+1} \ . \tag{6.36}$$

We remark that the bound (6.35) can be made to decrease as $c_n/n^{\frac{1}{2}+\delta}$ for arbitrarily small $\delta > 0$ by properly choosing the sequence $\{s_n\}$. This rate of decrease is thus arbitrarily close to the rate of decrease of the bound (6.30) for the deleted estimate.

6.4 Linear Discrimination Rules

A linear discrimination rule $\{\delta_n\}$ is a rule that is constructed as in (6.1) where

$$\psi_{nj}(V_n, x) = \sum_{i=0}^{m'} w_{ji}^n(V_n)\, \varphi_i(x) \ , 1 \le j \le M, \tag{6.37}$$

where $m' \ge 1$, $\varphi_0, \varphi_1, \ldots, \varphi_{m'}$ are Borel measurable mappings from $\mathbb{R}^m$ to $\mathbb{R}^1$ with $\varphi_0 \equiv 1$, and the $w_j^n = (w_{j0}^n, w_{j1}^n, \ldots, w_{jm'}^n), 1 \le j \le M$, are Borel measurable mappings from $(\mathbb{R}^m \times \{1, \ldots, M\})^n$ to $\mathbb{R}^{m'+1}$. We also require that the mappings ξ_n not depend upon x. Linear discrimination rules are thoroughly investigated in the literature on parametric discrimination (e.g., see Duda and Hart (1973) or Ho and Agrawala (1968) for surveys). The basic property of these rules is that for any data sequence V_n, the set for which $\psi_{nj}(V_n, x) > \psi_{nk}(V_n, x)$, $j \ne k$, is a linear halfspace if $\varphi_0, \varphi_1, \ldots, \varphi_{m'}$ are considered as the new variables.

We have seen that the resubstitution estimate is nearly useless to the statistician for most two-step rules. However we will show that L_n^R is a very good distribution-free estimate of L_n for linear discrimination rules. The qualification "very good" refers to the fact that it is possible to obtain an upper-bound for $P\{|L_n^R - L_n| \geq \epsilon\}$ that

(i) does not depend upon the distribution of (X, θ),

(ii) does not depend upon the way of choosing the w_{ji}^n, $1 \leq j \leq M, 0 \leq i \leq m'$, and

(iii) decreases exponentially fast with n.

The results of this section complement the results of Devroye and Wagner (1976). Using the bound of Vapnik and Chervonenkis (1971) it is possible to prove the following nontrivial result.

Theorem 6.8. Let $\{\delta_n\}$ be any linear discrimination rule, then, for all $\epsilon > 0$, all n and all distributions of (X, θ),

$$P\{|L_n^R - L_n| \geq \epsilon\} \leq 8M(1+n\epsilon/3M)^{2(m'+1)M} e^{-n\epsilon^3/144M^2 2^{2M}}.$$

No attempt has been made to obtain the best possible constants in theorem 6.8. The merit of the theorem is that the bound is valid for all ways of choosing the w_{ji}^n from the data, and all possible (unknown) distributions of (X, θ). Mimicking the proof of theorem 6.8, we can also prove that for $L_n^{R'}$ and any linear discrimination rule,

$$\sup_{(\pi_1, \ldots, \pi_M, \mu_1, \ldots, \mu_M)} P\{|L_n^{R'} - L_n| \geq \epsilon\}$$

$$\leq 6M(1+2n\epsilon/3M)^{2(m'+1)M} e^{-n\epsilon^3/18M^2 2^{2M}}.$$

6.5 Linear Ordering Rules and $\{k_n\}$-Local Rules

There is another large class of discrimination rules for which it is very easy to find useful distribution-free estimates of the conditional probability of error, that is, the class of $\{k_n\}$-local rules. First we will obtain distribution-free upper-bounds for $P\{|L_n^D - L_n| \geq \epsilon\}$ using techniques that were suggested by Cover (1969) and Rogers and Wagner (1976). Similar new upper-bounds are obtained with the holdout estimate. We will see that the resubstitution estimate fails as a distribution-free estimate of L_n for $\{k_n\}$-local rules in general but that L_n^R is a useful distribution-free estimate of L_n for $\{k_n\}$-nearest neighbor rules provided that k_n grows large. Let us first clearly define the classes of rules that will be discussed in this section.

Let us enlarge V_n in the following manner. Let $V_n' = (X_1, \theta_1, Z_1), \ldots, (X_n, \theta_n, Z_n)$ where $Z_1, \ldots, Z_n$ are random variables that are independent of V_n and have the property that equality between any two Z_i's occurs with probability zero. For example, we can let $Z_i = 1/i, i \geq 1$, or else, we can let $Z_1, \ldots, Z_n$ be independent all having a uniform distribution over $[0,1]$. Given $x \in \mathbb{R}^m$, the Z_i are used to obtain a wp1 uniquely defined permutation $(X_1^x, \theta_1^x, Z_1^x), \ldots, (X_n^x, \theta_n^x, Z_n^x)$ of V_n' with the property that $\|X_1^x - x\| \leq \ldots \leq \|X_n^x - x\|$ and if $\|X_i^x - x\| = \|X_{i+1}^x - x\|$, then $Z_i^x < Z_{i+1}^x$. Let $V_n^x = (X_1^x, \theta_1^x, Z_1^x), \ldots, (X_{k_n}^x, \theta_{k_n}^x, Z_{k_n}^x)$ where $1 \leq k_n \leq n$. We assume that V_n' replaces V_n in the definition of L_n and that a decision function δ_n is a Borel measurable function of V_n' and x. If δ_n is a Borel measurable function of V_n^x and x for all n, then we say that $\{\delta_n\}$ is a $\{k_n\}$-local rule where $1 \leq k_n \leq n$ for all n. For such rules it is very easy to define a close deleted decision function $\tilde{\delta}_n$. Let

V'_{ni} be the sequence that is obtained by deleting (X_i, θ_i, Z_i) from V'_n and obtain V^x_{ni} from V'_{ni} using the same procedure to get V^x_n from V'_n. Thus, both V^x_n and V^x_{ni} are random vectors taking values in $(\mathbb{R}^m \times \{1,\ldots,M\} \times \mathbb{R})^{k_n}$. If $1 \le k_n \le n-1$ and

$$\delta_n(V'_n, x) = g_n(V^x_n, x) \tag{6.38}$$

for some vector function g_n, then let the deleted decision function be defined by

$$\tilde{\delta}_n(V'_{ni}, x) = g_n(V^x_{ni}, x), 1 \le i \le n. \tag{6.39}$$

Similarly, if $1 \le s_n \le n$ and $1 \le k_n \le n-s_n$, and if $T'_n = (X_1, \theta_1, Z_1), \ldots, (X_{n-s_n}, \theta_{n-s_n}, Z_{n-s_n})$, we can obtain T^x_n from T'_n just as we obtained V^x_n from V'_n. If δ_n is defined by (6.38), then let the holdout decision function be

$$\tilde{\delta}_n(T'_n, x) = g_n(T^x_n, x).$$

The crucial observation in this section which will enable us to use the powerful theorems 6.1 and 6.2 is the following.

$$\begin{aligned} E\{|\delta_{n\theta}(V'_n, X) - \tilde{\delta}_{n\theta}(T'_n, X)|\} &\le E\{|g_{n\theta}(V^X_n, X) - g_{n\theta}(T^X_n, X)|\} \\ &\le P\{V^X_n \ne T^X_n\} \le \sup_x P\{V^x_n \ne T^x_n\} \\ &\le \sup_x P\{\bigcup_{i=1}^{s_n} \bigcup_{j=1}^{k_n} \{Z_{n-s_n+i} = Z^x_j\}\} \\ &\le s_n k_n / n \end{aligned} \tag{6.40}$$

for the holdout decision function $\tilde{\delta}_n$ when $\{\delta_n\}$ is a $\{k_n\}$-local rule. Similarly, for the deleted decision function $\tilde{\delta}_n$,

$$E\{|\delta_{n\theta}(V'_n, X) - \tilde{\delta}_{n\theta}(V'_{n1}, X)|\} \le k_n / n. \tag{6.41}$$

A simple combination of (6.40) and (6.41) with theorems 6.1 and 6.2 yields the following interesting results.

Theorem 6.9. If $\{\delta_n\}$ is any $\{k_n\}$-local rule, then for all $\epsilon>0$,

$$P\{|L_n^D - L_n| \geq \epsilon\} \leq (1+6k_n)/n\epsilon^2 \tag{6.42}$$

and

$$E\{(L_n^D - L_n)^2\} \leq (1+6k_n)/n \tag{6.43}$$

for all distributions of (X,θ) provided that $1 \leq k_n \leq n-1$.

Theorem 6.10. If $\{\delta_n\}$ is any $\{k_n\}$-local rule and if $1 \leq s_n \leq n-k_n$, then for all $\epsilon>0$,

$$P\{|L_n^H - L_n| \geq \epsilon\} \leq 2e^{-s_n\epsilon^2/2} + 2s_n k_n/n\epsilon \tag{6.44}$$

and

$$E\{(L_n^H - L_n)^2\} \leq 1/2s_n + 2s_n k_n/n \tag{6.45}$$

for all distributions of (X,θ).

Several remarks are in order. The bounds (6.42) and (6.43) tend to 0 as $n\to\infty$ if $k_n/n \overset{n}{\to} 0$. Very interestingly, it is always possible to find a sequence $\{s_n\}$ for the holdout estimate such that the bounds (6.44) and (6.45) tend to 0 as $n\to\infty$ provided that $k_n/n \overset{n}{\to} 0$ (just let $s_n \sim \sqrt{n/k_n}$). One rule to which the above mentioned theorems apply is the generalized nearest neighbor rule of section 5.2 where the sequence of weight vectors $v_n=(v_{n1},\ldots,v_{nn})$ satisfies

$$\sum_{i=k_n+1}^{n} v_{ni} = 0 \ , n\geq 1. \tag{6.46}$$

A special class of $\{k_n\}$-local rules is the class of $\{k_n\}$-nearest neighbor rules. For these rules we can considerably improve (6.42)-(6.45).

A $\{k_n\}$-nearest neighbor rule $\{\delta_n\}$ is a sequence of decision functions constructed as (6.1) where

$$\psi_{nj}(V'_n, x) = \sum_{i=1}^{k_n} I_{\{\theta_i^x = j\}}, \quad 1 \le j \le M. \tag{6.47}$$

Thus, ψ_{nj} counts the number of $(X_i^x, \theta_i^x, Z_i^x)$'s in V_n^x for which $\theta_i^x = j$. We will let the deleted decision function be

$$\tilde{\delta}_n(V'_{ni}, x) = \xi_n(J', x), \quad 1 \le i \le n, \tag{6.48}$$

where the J'-th subset of $\{1, \ldots, M\}$ is the subset of indices for which $\psi'_{nj} = \underset{1 \le i \le M}{\text{Max}} \psi'_{ni}$, where

$$\psi'_{nj}(V'_{ni}, x) = \sum_{i=1}^{\text{Min}(k_n, n-1)} I_{\{\theta_i^x = j\}}, \quad 1 \le j \le M, 1 \le i \le n, \tag{6.49}$$

and where $(X_1^x, \theta_1^x, Z_1^x), \ldots, (X_{n-1}^x, \theta_{n-1}^x, Z_{n-1}^x)$ is the ordered permutation of V'_{ni}. With this definition we can define a deleted decision function for an n-nearest neighbor rule as well. A similar definition can be given for the holdout decision function. The main result for $\{k_n\}$-nearest neighbor rules is that both the deleted and the holdout estimate are useful tools for the statistician to estimate L_n. Indeed, it is not very hard to prove the following inequalities, all valid for M=2.

Theorem 6.11. Let $\{\delta_n\}$ be a $\{k_n\}$-nearest neighbor rule and let $\epsilon > 0$ be arbitrary. Then, if $1 \le k_n \le n-1$,

$$P\{|L_n^D - L_n| \ge \epsilon\} \le (1 + 24\sqrt{k_n/\pi})/n\epsilon^2$$

and

$$E\{(L_n^D - L_n)^2\} \le (1 + 24\sqrt{k_n/\pi})/n$$

for all distributions of (X, θ).

Theorem 6.12. Let $\{\delta_n\}$ be a $\{k_n\}$-nearest neighbor rule and let $\epsilon>0$ be arbitrary. If $1 \le k_n \le n-s_n$, then

$$P\{|L_n^H - L_n| \ge \epsilon\} \le 2e^{-s_n\epsilon^2/2} + (8/\sqrt{\pi})s_n\sqrt{k_n}/n\epsilon$$

and

$$E\{(L_n^H - L_n)^2\} \le 1/2s_n + (8/\sqrt{\pi})\, s_n\sqrt{k_n}/n$$

for all distributions of (X,θ).

Notice that $k_n \le n$ so that

$$E\{(L_n^D - L_n)^2\} \le 1/n + 24/\sqrt{\pi n} \tag{6.50}$$

for all the $\{k_n\}$-nearest neighbor rules and for all the distributions of (X,θ), independent of k_n. The rate of decrease of this bound is $1/\sqrt{n}$. We can show that over all the $\{k_n\}$-nearest neighbor rules, this rate is actually the best possible rate of decrease.

Theorem 6.13. If $M=2$, n is even and $0<\epsilon<1/2$, then there exists a $\{k_n\}$-nearest nearest neighbor rule for some sequence $\{k_n\}$ such that

$$\sup_{(\pi_1,\ldots,\pi_M,\mu_1,\ldots,\mu_M)} E\{(L_n^D - L_n)^2\} \ge 1/4\sqrt{2n}$$

and

$$\sup_{(\pi_1,\ldots,\pi_M,\mu_1,\ldots,\mu_M)} P\{|L_n^D - L_n| \ge \epsilon\} \ge 1/\sqrt{2n}\,.$$

The resubstitution estimate L_n^R is not useful with the nearest neighbor rule ($k_n=1$ for all n) because in that case $L_n^R=0$. However, if $M=2$, $\pi_1=\pi_2=\frac{1}{2}$ and if μ_1 and μ_2 are uniform measures over $[0,1]$, then $L_n=\frac{1}{2}$ for all n. Thus, $L_n - L_n^R=\frac{1}{2}$ for some distribution of (X,θ) for all n. We state this as a theorem.

Theorem 6.14. If $\{\delta_n\}$ is a nearest neighbor rule and M=2, then there exists a distribution of (X,θ) such that $L_n=\frac{1}{2}$ and $L_n^R=0$ for all n.

However, for $\{k_n\}$-nearest neighbor rules, the resubstitution estimate is useful if k_n grows large. We could of course expect this because the deleted and the resubstitution estimate are close to each other for large k_n. The following result should be compared with the theorems 6.9 and 6.12 for deleted estimates.

Theorem 6.15. Let M=2 and let $\{\delta_n\}$ be a $\{k_n\}$-nearest neighbor rule. Then, for all the distributions of (X,θ),

$$E\{(L_n^R-L_n)^2\} \le 2\,(1+24\sqrt{k_n/\pi})/n + 8/\sqrt{\pi k_n}$$

and

$$E\{(L_n^R-L_n)^2\} \le 2\,(1+6k_n)/n + 8/\sqrt{\pi k_n}\,.$$

The $\{k_n\}$-nearest neighbor rules are but a special case of linear ordering rules, that is, rules defined by (6.1) with

$$\psi_{nj}(V_n',x) = \sum_{i=1}^{n} w_{ji}^n(X_i^x,\theta_i^x,x)\,I_{\{\theta_i^x=j\}},\ 1\le j\le M, \qquad (6.51)$$

where the w_{ji}^n are real-valued Borel measurable functions on $\mathbb{R}^m\times\{1,\ldots,M\}\times\mathbb{R}$, $1\le j\le M$, $1\le i\le n$, $n\ge 1$. For $\{k_n\}$-nearest neighbor rules, it is clear that $w_{ji}^n=1$ if $1\le i\le k_n$ and $w_{ji}^n=0$ otherwise. The powerful rule of section 5.2 is an example of a linear ordering rule with $w_{ji}^n=v_{ni}$, $1\le i\le n$. We have seen that $P\{|L_n^R-L_n|\ge\epsilon\}$ can not be upper bounded by a function of n that decreases to 0 as $n\to\infty$ uniformly over all distributions of (X,θ) and over all the sequences of weight vectors $\{v_n\}$ (just let $v_{n1}=1, v_{ni}=0, i>1$). However, it remains an open question whether such a bound can be found with L_n^D. One property that may be exploited in trying to

find such a bound is that given $X_1^x, \ldots, X_n^x$, the $w_{ji}^n(X_i^x, \theta_i^x, x)\, I_{\{\theta_i^x = j\}}$, $1 \le i \le n$, are independent random variables.

6.6 Proofs

Proof of theorem 6.1

Let $(X_0, \theta_0), (X_1, \theta_1), \ldots, (X_n, \theta_n), (X, \theta)$ be iid random vectors distributed as (X, θ). Note first that

$$E\{L_n^2\} = E\{\left(P\{\theta_{V_n, X} \neq \theta \mid V_n\}\right)^2\}$$

$$= E\{P\{\theta_{V_n, X_0} \neq \theta_0 \; ; \; \theta_{V_n, X} \neq \theta \mid V_n\}\}$$

$$= E\{(1-\delta_{n\theta_0}(V_n, X_0))(1-\delta_{n\theta}(V_n, X))\},$$

$$E\{L_n^D L_n\} = E\{(1-\delta_{n\theta}(V_n, X))(1-\tilde{\delta}_{n\theta_1}(V_{n1}, X_1))\},$$

and

$$E\{(L_n^D)^2\} = E\{n^{-2}\sum_{i=1}^{n} I_{\{\theta_{V_{ni}}, X_i \neq \theta_i\}}\}$$

$$+ n^{-2} E\{\sum_{i \neq j} I_{\{\theta_{V_{ni}}, X_i \neq \theta_i\}} I_{\{\theta_{V_{nj}}, X_j \neq \theta_j\}}\}$$

$$= n^{-1} E\{1-\tilde{\delta}_{n\theta_1}(V_{n1}, X_1)\}$$

$$+ (1-1/n) E\{(1-\tilde{\delta}_{n\theta_1}(V_{n1}, X_1))(1-\tilde{\delta}_{n\theta_2}(V_{n2}, X_2))\}.$$

With $\delta_n' = 1 - \delta_n$, $\tilde{\delta}_n' = 1 - \tilde{\delta}_n$, we obtain

$$E\{(L_n^D - L_n)^2\} = E\{\delta'_{n\theta_0}(V_n, X_0)\, \delta'_{n\theta}(V_n, X) + \tilde{\delta}'_{n\theta_1}(V_{n1}, X_1)\, \tilde{\delta}'_{n\theta_2}(V_{n2}, X_2)$$

$$- 2\, \delta'_{n\theta}(V_n, X)\, \tilde{\delta}'_{n\theta_1}(V_{n1}, X_1)\}$$

$$+ n^{-1} E\{\tilde{\delta}'_{n\theta_1}(V_{n1}, X_1) - \tilde{\delta}'_{n\theta_1}(V_{n1}, X_1)\, \tilde{\delta}'_{n\theta_2}(V_{n2}, X_2)\}.$$

The last term is clearly upper-bounded by $1/n$. Now, let $a, b \in \{0,1,2,c\}$, let $(X_c, \theta_c)=(X,\theta)$, let $V_{ab}=(X_a,\theta_a),(X_b,\theta_b),(X_3,\theta_3),\ldots,(X_n,\theta_n)$ and let $V_a=(X_a,\theta_a),(X_3,\theta_3),\ldots,(X_n,\theta_n)$. With this notation, we have

$$E\{(L_n^D-L_n)^2\} \le 1/n + E\{\delta'_{n\theta_0}(V_{12},X_0)\,\delta'_{n\theta}(V_{12},X) + \tilde{\delta}'_{n\theta_1}(V_2,X_1)\,\tilde{\delta}'_{n\theta_2}(V_1,X_2) - 2\,\delta'_{n\theta}(V_{12},X)\,\tilde{\delta}'_{n\theta_1}(V_2,X_1)\}$$

and

$$\begin{aligned}
&E\{\delta'_{n\theta_0}(V_{12},X_0)\,\delta'_{n\theta}(V_{12},X) - \delta'_{n\theta}(V_{12},X)\,\tilde{\delta}'_{n\theta_1}(V_2,X_1)\} \\
&= E\{\delta'_{n\theta_1}(V_{0c},X_1)\,\delta'_{n\theta_2}(V_{0c},X_2) - \delta'_{n\theta_1}(V_{02},X_1)\,\tilde{\delta}'_{n\theta_2}(V_0,X_2)\} \\
&= E\{\delta'_{n\theta_1}(V_{0c},X_1)\,\delta'_{n\theta_2}(V_{0c},X_2) - \tilde{\delta}'_{n\theta_1}(V_0,X_1)\,\delta'_{n\theta_2}(V_{0c},X_2) \\
&\quad + \tilde{\delta}'_{n\theta_1}(V_0,X_1)\,\delta'_{n\theta_2}(V_{0c},X_2) - \delta'_{n\theta_1}(V_{02},X_1)\,\delta'_{n\theta_2}(V_{0c},X_2) \\
&\quad + \delta'_{n\theta_1}(V_{02},X_1)\,\delta'_{n\theta_2}(V_{0c},X_2) - \delta'_{n\theta_1}(V_{02},X_1)\,\tilde{\delta}'_{n\theta_2}(V_0,X_2)\} \\
&\le 3E\{|\delta'_{n\theta}(V_{12},X) - \tilde{\delta}'_{n\theta}(V_2,X)|\} \\
&= 3E\{|\delta_{n\theta}(V_{12},X) - \tilde{\delta}_{n\theta}(V_2,X)|\}\,.
\end{aligned}$$

Further, in a similar fashion,

$$\begin{aligned}
&E\{\tilde{\delta}'_{n\theta_1}(V_2,X_1)\,\tilde{\delta}'_{n\theta_2}(V_1,X_2) - \delta'_{n\theta}(V_{12},X)\,\tilde{\delta}'_{n\theta_1}(V_2,X_1)\} \\
&= E\{\tilde{\delta}'_{n\theta_1}(V_2,X_1)\,\tilde{\delta}'_{n\theta_2}(V_1,X_2) - \delta'_{n\theta_1}(V_{c2},X_1)\,\tilde{\delta}'_{n\theta_2}(V_c,X_2)\} \\
&= E\{\tilde{\delta}'_{n\theta_1}(V_2,X_1)\,\tilde{\delta}'_{n\theta_2}(V_1,X_2) - \tilde{\delta}'_{n\theta_1}(V_2,X_1)\,\delta'_{n\theta_2}(V_{c1},X_2) \\
&\quad + \tilde{\delta}'_{n\theta_1}(V_2,X_1)\,\delta'_{n\theta_2}(V_{c1},X_2) - \delta'_{n\theta_1}(V_{c2},X_1)\,\delta'_{n\theta_2}(V_{c1},X_2) \\
&\quad + \delta'_{n\theta_1}(V_{c2},X_1)\,\delta'_{n\theta_2}(V_{c1},X_2) - \delta'_{n\theta_1}(V_{c2},X_1)\,\tilde{\delta}'_{n\theta_2}(V_c,X_2)\}
\end{aligned}$$

$$\leq 3\, E\{|\delta_{n\theta}(V_{12},X)-\tilde{\delta}_{n\theta}(V_2,X)|\}\,.$$

This concludes the proof of theorem 6.1.

Q.E.D.

Proof of theorem 6.2

By

$$|L_n-L_n^H| \leq |L_n-E\{L_n^H|T_n\}| + |E\{L_n^H|T_n\}-L_n^H|\,,$$

$$|L_n-E\{L_n^H|T_n\}| = |P\{\theta_{V_n,X}\neq\theta|V_n\}-P\{\theta_{T_n,X}\neq\theta|T_n\}|$$

$$\leq E\{|\delta_{n\theta}(V_n,X)-\tilde{\delta}_{n\theta}(T_n,X)|\,|V_n\}\,,$$

$$E\{\left(E\{L_n^H|T_n\}-L_n^H\right)^2|T_n\} \leq 1/4s_n\,,$$

the c_r-inequality and the Cauchy inequality for conditional expectations, we have,

$$E\{(L_n-L_n^H)^2\} \leq 2\left(E\{(L_n-E\{L_n^H|T_n\})^2\}+E\{(L_n^H-E\{L_n^H|T_n\})^2\}\right)$$

$$\leq 2\,E\{\left(E\{|\delta_{n\theta}(V_n,X)-\tilde{\delta}_{n\theta}(T_n,X)|\,|V_n\}\right)^2\} + 1/2s_n$$

$$\leq 2\,E\{|\delta_{n\theta}(V_n,X)-\tilde{\delta}_{n\theta}(T_n,X)|^2\} + 1/2s_n\,.$$

Next,

$$P\{|E\{L_n^H|T_n\}-L_n^H| \geq \epsilon\,|T_n\} \leq 2e^{-2s_n\epsilon^2} \text{ wpl } 1\,,\ \epsilon>0,$$

implies that

$$P\{|L_n-L_n^H| \geq \epsilon\}$$

$$\leq P\{|L_n-E\{L_n^H|T_n\}| \geq \epsilon/2\} + P\{|L_n^H-E\{L_n^H|T_n\}| \geq \epsilon/2\}$$

$$\leq 2e^{-s_n\epsilon^2/2} + (4/\epsilon^2)\,E\{\left(E\{|\delta_{n\theta}(V_n,X)-\tilde{\delta}_{n\theta}(T_n,X)|\,|V_n\}\right)^2\}$$

$$\leq 2\, e^{-s_n \epsilon^2/2} + E\{|\delta_{n\theta}(V_n,X) - \delta_{n\theta}(T_n,X)|^2\}.$$

Finally, (6.23) is proved as (6.22).

Q.E.D.

Proof of theorem 6.3

Let M=2 and let

$$w_j^n(x,y) = I_{\{\|x-y\| \leq \frac{1}{2}\}}, \quad 1 \leq j \leq M, x, y \in \mathbb{R}^m .$$

Then

$$\delta_n(V_n,x) = \begin{cases} (1,0) & \text{if } \psi_{n1}(V_n,x) > \psi_{n2}(V_n,x) \\ (0,1) & \text{if } \psi_{n1}(V_n,x) < \psi_{n2}(V_n,x) \\ \xi_n(3,x) & \text{if } \psi_{n1}(V_n,x) = \psi_{n2}(V_n,x) \end{cases}$$

where for every x, $\xi_n(3,x)$ is an arbitrary probability vector. Let μ put mass $1/2n$ at $1,2,\ldots,2n$ and let μ_1 put mass 0 at those x's for which $\xi_{n1}(3,x) \geq \frac{1}{2}$. Let μ_2 put mass 0 at those x's for which $\xi_{n1}(3,x) < \frac{1}{2}$. Then, if N_{jn}^k is the number of X_i's with $X_i = k$ and $\theta_i = j$, we have

$$L_n = \frac{1}{2n} \sum_{k=1}^{2n} I_{\{N_{1n}^k = N_{2n}^k = 0\}} \operatorname{Max}(\xi_{n1}(3,k), \xi_{n2}(3,k))$$

$$\geq \tfrac{1}{2} n / 2n \geq \tfrac{1}{4}$$

and

$$L_n^R = 0.$$

The same proof carries through if

$$w_j^n(x,y) = \sum_{k=1}^{\infty} I_{\{x \in B_k\}} I_{\{y \in B_k\}}, \quad 1 \leq j \leq M, x, y \in \mathbb{R}^m,$$

where $\{B_1, B_2, \ldots\}$ is any countable partition of $\mathbb{R}^m$, provided that $\pi_1 = 1, \pi_2 = 0$ and μ_1 puts mass $1/2n$ in $B_1, \ldots, B_{2n}$ and mass 0

outside $\bigcup_{i=1}^{2n} B_i$.

Q.E.D.

Proof of theorem 6.4

Assume without loss of generality that $n \geq 3$ and let $p=1/n$ and pick a number a from $(1/(n-1), 1/(n-2))$. Construct a two-step rule with

$$w_j^n(y,x) = \begin{cases} I_{\{\|y-x\| \leq 1\}} & ,j=1, \\ a\, I_{\{\|y-x\| \leq 1\}} & ,j=2, \end{cases} \qquad ,x,y \in R^m .$$

Let $\pi_1=p, \pi_2=1-p$ and let μ_1 and μ_2 both put mass 1 at 0. Let N_{jn} be the number of X_i's with $\theta_i=j$, $j=1,2$. Notice that $N_{1n}+N_{2n}=1$ and that the equality $N_{1n}=N_{2n}a$ cannot occur by the choice of a. Since $X=X_1=X_2=\ldots=X_n=0$ wp1, we see that wp1,

$$\theta_{V_n,X} = \begin{cases} 1 & \text{if } N_{1n}>N_{2n}a \\ 0 & \text{if } N_{1n}<N_{2n}a \end{cases} .$$

If $N_{1n}=1$ and $N_{2n}=n-1$, then $L_n=p$ in view of $(n-1)a>1$. Because $(n-2)a < 1$, it is obvious that in that case, $L_n^D=(n-1)/n$. Thus,

$$P\{|L_n-L_n^D| \geq (n-1)/n - p\} = P\{|L_n-L_n^D| \geq (n-2)/n\}$$

$$\geq P\{N_{1n}=1\} = \binom{n}{1} p(1-p)^{n-1} \geq np(1-p)^n = (1-p)^n$$

$$\geq e^{-np/(1-p)} > e^{-2}$$

because $p<\frac{1}{2}$. Inequality (6.29) follows from this and $(n-2)/n \geq \frac{1}{3}$. Further,

$$E\{(L_n-L_n^D)^2\} \geq P\{|L_n-L_n^D| \geq \tfrac{1}{3}\}/9 > 1/9e^2 .$$

Notice that the same is true if

$$w_j^n(y,x) = 1/(1+\|y\|)^{\ell}\ ,j=1,2\ ;\ x,y\in\mathbb{R}^m,$$

where $\ell>0$, provided that $\pi_1=p, \pi_2=1-p$, and that μ_1 and μ_2 put mass 1 at 0 and z respectively, where z is chosen such that $1/(1+\|z\|)^{\ell}=a$.

Q.E.D.

Proofs of theorems 6.5 and 6.7

The mathematical machinery needed to prove these theorems is rather heavy. To see what is going on, we extracted the following lemmas that can be of separate interest to the reader. The lemmas are special versions of the Berry-Esseen central limit theorem (Feller, 1966 ; Osipov and Petrov, 1967 ; Hertz, 1969).

Lemma 6.1. If $a,b\in\mathbb{R}$ with $a<b$, if $Y_1,\ldots,Y_n$ are independent identically distributed random variables with variance $\sigma^2<\infty$ and $|Y_1-E\{Y_1\}|\leq g<\infty$ wp1, then

$$P\{a\leq\left(\sum_{i=1}^{n} Y_i\right)/\sigma\sqrt{n}\ \leq b\}\leq (b-a)/\sqrt{2\pi}+2C_0 g/\sigma\sqrt{n}$$

where

$$C_0 = 1 + 14/3\sqrt{2\pi} + 64\,C_1$$

and C_1 is the universal constant in the Berry-Esseen theorem (e.g., according to Zolotarev (1966), $C_1<1.322$).

Proof of lemma 6.1

Let

$$\Phi(x) = (1/\sqrt{2\pi})\int_{-\infty}^{x} e^{-t^2/2}dt\ ,x\in\mathbb{R}.$$

From Hertz (1969) we know that

$$\sup_x\ |P\{\left(\sum_{i=1}^{n}\left(Y_i-E\{Y_i\}\right)\right)/\sigma\sqrt{n}\ \}-\Phi(x)|$$

$$\leq \left(C_0/(\sigma\sqrt{n})^3\right) \int_0^{\sigma\sqrt{n}} n \int_{\{x:|x|>u\}} x^2 dF(x)\, du$$

where F is the distribution function of $Y_1 - E\{Y_1\}$. We bound the last expression by

$$C_0 n g \sigma^2/\sigma^3 n^{3/2} \leq C_0 g/\sigma\sqrt{n} \quad ,$$

and lemma 6.1 follows from this and $\Phi(b)-\Phi(a) \leq (b-a)/\sqrt{2\pi}$.

Q.E.D.

Lemma 6.2. If $Y_1, \ldots, Y_n, Z_1, \ldots, Z_s$ are independent identically distributed random variables taking values in $[-1,-c] \cup \{0\} \cup [c,1]$ where $0<c\leq 1$, then

$$P\{ \operatorname{sgn}(\sum_{i=1}^{n} Y_i + \sum_{j=1}^{s} Z_j) \neq \operatorname{sgn}(\sum_{i=1}^{n} Y_i) \} \leq s\, C_2/c\sqrt{n+1}$$

where

$$C_2 = 4/\sqrt{\pi} + 4C_0\sqrt{2}$$

and where sgn(.) is the sign function ,

$$\operatorname{sgn}(u) = \begin{cases} 1 & \text{if } u>0 \\ 0 & \text{if } u<0 \\ \frac{1}{2} & \text{if } u=0 \, . \end{cases}$$

Proof of lemma 6.2

Let $Y = \sum_{i=1}^{n} I_{\{Y_i \neq 0\}}$, $Z = \sum_{i=1}^{s} I_{\{Z_i \neq 0\}}$, $S_Y = \sum_{i=1}^{n} Y_i$ and $S_Z = \sum_{i=1}^{s} Z_i$. Let A denote the event $\{\operatorname{sgn}(S_Y+S_Z) \neq \operatorname{sgn}(S_Y)\}$. Obviously,

$$P\{A\} \leq \sum_{z=1}^{s} \sum_{y=0}^{n} P\{Y=y\} P\{Z=z\} P\{A \mid Y=y, Z=z\} .$$

Let $\lambda = E\{Y_1\} = E\{Y_1 I_{\{Y_1 \neq 0\}}\} = p\, E\{Y_1 \mid Y_1 \neq 0\}$ where $p=P\{Y_1 \neq 0\}$. If $\lambda^2 < p^2 c^2/2$, then

$$\sigma_0^2 = E\{(Y_1 - E\{Y_1 \,|\, Y_1 \neq 0\})^2 \,|\, Y_1 \neq 0\}$$
$$= E\{Y_1^2 \,|\, Y_1 \neq 0\} - (E\{Y_1 \,|\, Y_1 \neq 0\})^2 \geq c^2 - c^2/2 = c^2/2.$$

In that case, by lemma 6.1,

$$P\{A \,|\, Y=y, Z=z\} \leq P\{\,|S_Y| \leq z \,|\, Y=y\}$$
$$\leq \text{Min}\left(1 \,;\, \left(2C_0 + 2z/\sqrt{2\pi}\right)/(\sigma_0\sqrt{y})\right)$$
$$\leq \text{Min}\left(1 \,;\, \left(2C_0\sqrt{2} + 2z/\sqrt{\pi}\right)/(c\sqrt{y})\right) .$$

If $\lambda^2 \geq p^2c^2/2$, then, by Chebyshev's inequality, if $q = \lambda/p = E\{Y_1 \,|\, Y_1 \neq 0\}$,

$$P\{A \,|\, Y=y, Z=z\} \leq P\{S_Y + S_Z < 0, S_Y \geq 0 \,|\, Y=y, Z=z\}$$
$$+ P\{S_Y + S_Z \geq 0, S_Y < 0 \,|\, Y=y, Z=z\}$$
$$\leq P\{S_Y + S_Z - (y+z)q < -(y+z)q \,|\, Y=y, Z=z\}$$
$$+ P\{S_Y - yq < -yq \,|\, Y=y, Z=z\}$$
$$\leq 2\sigma_0^2/yq^2 \leq 4/yc^2.$$

Note that if $4/yc^2 \leq 1$, then $4/yc^2 \leq 2/c\sqrt{y} \leq 2\sqrt{2}\, C_0/c\sqrt{y}$ so that, since $z \geq 1$ in the summation,

$$P\{A \,|\, Y=y, Z=z\} \leq z\, \text{Min}\left(1 \,,\, C_3/2c\sqrt{y}\right)$$

where $C_3 = 2/\sqrt{\pi} + 2C_0\sqrt{2} \triangleq C_2/2$. Noting that $\sum_{z=1}^{s} z\, P\{Z=z\} = sp$, we have, when L is a binomial random variable with parameters $(n+1)$ and p,

$$P\{A\} \leq \sum_{y=0}^{n} s \binom{n}{y} p^{y+1}(1-p)^{n-y} \text{Min}\left(1 \,;\, C_3/c\sqrt{y}\right)$$
$$= s \sum_{y=0}^{n} \frac{y+1}{n+1} \binom{n+1}{y+1} p^{y+1}(1-p)^{n-y}\, \text{Min}\left(1 \,;\, C_3/c\sqrt{y}\right)$$
$$\leq (s/(n+1))\, E\left\{\text{Min}\left(L+1 \,;\, C_3(L+1)/c\sqrt{L}\right)\right\}$$

$$\leq (s/(n+1))\ E\left\{ \mathrm{Min}\left(L+1\ ;\ (\sqrt{L}+1)C_3/c \right) \right\}$$

$$\leq (s/(n+1))\ \mathrm{Min}\left(E\{L\}+1\ ;\ (\sqrt{E\{L\}}+1)C_3/c \right)$$

$$\leq (sC_3/c(n+1))\left(1+\sqrt{(n+1)p} \right) \leq 2sC_3/c\sqrt{n+1}$$

where we used Jensen's inequality, $1\leq\sqrt{n+1}$ and $p\leq 1$.

Q.E.D.

Lemma 6.3. Let $\{\delta_n\}$ be a two-step rule, let $\tilde{\delta}_n$ be the holdout decision function corresponding to δ_n and let δ_n have ratio range c_n. Then

$$\sup_{(\pi_1,\ldots,\pi_M,\mu_1,\ldots,\mu_M)} \sup_{\substack{x\in R^m \\ 1\leq j\leq M}} E\{ |\delta_{nj}(V_n,x)-\tilde{\delta}_{nj}(T_n,x)| \}$$

$$\leq C_2(M-1)c_n s_n/\sqrt{n-s_n+1}\ .$$

If $\tilde{\delta}_n$ is the deleted decision function, then the inequality holds with $T_n=V_{n1}$ and $s_n=1$.

Proof of lemma 6.3

Let $\pi_1,\ldots,\pi_M,\mu_1,\ldots,\mu_M$ be fixed and let $x\in R^m$. Consider first the holdout decision function $\tilde{\delta}_n$. The proof for the deleted decision function is similar.

$$E\{|\delta_{n1}(V_n,x)-\tilde{\delta}_{n1}(T_n,x)|\} \leq P\{\delta_{n1}(V_n,x)\neq\tilde{\delta}_{n1}(T_n,x)\}$$

$$\leq P\{\bigcup_{j=2}^{M}\{\ \mathrm{sgn}(\sum_{i=1}^{n} Y_i^{(j)}) \neq \mathrm{sgn}(\sum_{i=1}^{n-s_n} Y_i^{(j)})\ \}\}$$

where

$$Y_i^{(j)}=I_{\{\theta_i=1\}}w_1^n(X_i,x) - I_{\{\theta_i=j\}}w_j^n(X_i,x)\ ,\ 1\leq i\leq n, 2\leq j\leq M.$$

We used the fact that by the definition of ξ_n, Ψ_n and Ψ_n',

$$\bigcap_{j=2}^{M} \left| \{\operatorname{sgn}(\sum_{i=1}^{n} Y_i^{(j)}) = \operatorname{sgn}(\sum_{i=1}^{n-s_n} Y_i^{(j)})\} \cap \{\delta_{n1}(V_n,x) \neq \tilde{\delta}_{n1}(T_n,x)\} \right|$$

is empty. Since $Y_1^{(j)}, \ldots, Y_n^{(j)}$ are independent, identically distributed random variables taking values in $\{0\} \cup [1, c_n] \cup [-c_n, -1]$ (without loss of generality), we thus have by lemma 6.2,

$$E\{|\delta_{n1}(V_n,x) - \tilde{\delta}_{n1}(T_n,x)|\} \leq (M-1)C_2 s_n c_n / \sqrt{n-s_n+1} .$$

Q.E.D.

To show theorem 6.7, we use (6.21) and lemma 6.3 as follows.

$$\begin{aligned} E\{(L_n - L_n^H)^2\} &\leq 1/2s_n + 2E\{|\delta_{n\theta}(V_n,X) - \tilde{\delta}_{n\theta}(T_n,X)|\} \\ &\leq 1/2s_n + 2 \sup_{\substack{x \in \mathbb{R}^m \\ 1 \leq j \leq M}} E\{|\delta_{nj}(V_n,x) - \tilde{\delta}_{nj}(T_n,x)|\} \\ &\leq 1/2s_n + 2C_2(M-1)c_n s_n / \sqrt{n-s_n+1} . \end{aligned}$$

This bound is uniform over all distributions of (X, θ). Similarly, (6.35) follows from (6.23) and lemma 6.3. Because δ_n and the deleted decision function $\tilde{\delta}_n$ are symmetric, theorem 6.1 and inequality (6.15) can be used in combination with lemma 6.3 (where $s_n = 1$ and $T_n = V_{n1}$) to prove (6.30) and (6.31).

Q.E.D.

Proof of theorem 6.6

Let M=2 and let n be an even positive integer. Let $\pi_1 = \pi_2 = \frac{1}{2}$, let $\mu_1 = \mu_2$ and let $0 < \epsilon < \frac{1}{2}$. It is obvious that with any δ_n, $L_n = \frac{1}{2}$. Consider for instance the decision function δ_n with ratio range $c_n = 1$ that is constructed as follows. Let

$$w_j^n(y,x) = 1 \quad , y, x \in \mathbb{R}^m; 1 \leq j \leq M,$$

and let N_{1n} be the number of observations for which $\theta_i=1$. A two-step rule that uses these $w_j^n, j=1,2$, is defined by

$$\delta_{n1}(V_n,x) = \begin{cases} 1 & , \text{ if } N_{1n}>n/2 \\ \frac{1}{2} & , \text{ if } N_{1n}=n/2 \\ 0 & , \text{ if } N_{1n}<n/2 \ . \end{cases}$$

The deleted decision function, constructed as outlined in section 6.3, is such that $L_n^D=1$ if $N_{1n}=n/2$. Thus,

$$E\{(L_n-L_n^D)^2\} \geq (1/2)^2 P\{N_{1n}=n/2\} = \tfrac{1}{4}\binom{n}{n/2} 2^{-n} > 1/4\sqrt{2n}$$

where we used an inequality of Mitrinovic (1964). Further,

$$P\{|L_n^D-L_n| \geq \epsilon\} \geq P\{N_{1n}=n/2\} > 1/\sqrt{2n} \ .$$

The same results are obtained with

$$w_j^n(y,x) = I_{\{\|y-x\| \leq h_n\}} \ , y,x \in \mathbb{R}^m, 1 \leq j \leq M,$$

where $h_n>0$, provided that $\mu_1=\mu_2$ both put all their mass in a set S which has a diameter that is smaller than h_n.

Q.E.D.

Proof of theorem 6.8

Let $\pi_1,\ldots,\pi_M,\mu_1,\ldots,\mu_M$ be fixed. Notice that for ali $\epsilon>0$,

$$P\{|L_n^R-L_n| \geq \epsilon\} \leq P\{|L_n^R-L_n^{R'}| \geq \epsilon/2\} + P\{|L_n^{R'}-L_n| \geq \epsilon/2\}$$

$$\leq 2e^{-n\epsilon^2/2} + P\{|L_n^{R'}-L_n| \geq \epsilon/2\}.$$

We will upper bound $P\{|L_n^{R'}-L_n| \geq \epsilon/2\}$. Let

$$L_n^j = P\{\theta_{V_n, X} \neq j \mid V_n\} = E\{1-\delta_{nj}(V_n,X) \mid V_n\}, 1 \leq j \leq M,$$

where X has probability measure μ_j in $\mathbb{R}^m$. Let

$$N_{jn} = \sum_{i=1}^{n} I_{\{\theta_i = j\}}, \quad 1 \le j \le M,$$

be the sample sizes of the M states. Let further for $1 \le j \le M$,

$$L_n^{R'j} = \sum_{i=1}^{n} I_{\{\theta_i = j\}} (1 - \delta_{nj}(V_n, X_i)) \; / N_{jn}$$

and

$$\pi_{jn} = N_{jn}/n .$$

It should be clear that

$$L_n^{R'} - L_n = \sum_{j=1}^{M} (\pi_{jn} L_n^{R'j} - \pi_j L_n^j) .$$

If $a > 0$ and $1 - 2a > 0$, then

$$P\{|L_n - L_n^{R'}| \ge \epsilon/2\} \le P\{\bigcup_{j=1}^{M} \{|\pi_{jn} - \pi_j| \ge a\epsilon/2M\}\}$$

$$+ P\{\bigcap_{j=1}^{M} \{|\pi_{jn} - \pi_j| < a\epsilon/2M\}, |L_n^{R'} - L_n| \ge \epsilon/2\}$$

$$\le 2M e^{-na^2\epsilon^2/2M^2}$$

$$+ P\{\bigcap_{j=1}^{M} \{|\pi_{jn} - \pi_j| < a\epsilon/2M\}, |\sum_{j=1}^{M} \pi_j (L_n^{R'j} - L_n^j)| \ge (1-a)\epsilon/2\}$$

$$\le 2M e^{-na^2\epsilon^2/2M^2}$$

$$+ \sum_{j=1}^{M} P\{|\pi_{jn} - \pi_j| < a\epsilon/2M \; , \; |L_n^{R'j} - L_n^j| \ge (1-a)\epsilon/2M\pi_j\}$$

$$\le 2M e^{-na^2\epsilon^2/2M^2}$$

$$+ \sum_{j=1}^{M} P\{N_{jn} \ge n\epsilon(1-2a)/2M \; , \; |L_n^{R'j} - L_n^j| \ge (1-a)\epsilon/2M\pi_j\}$$

where we used Hoeffding's inequality (1963) and the fact that we can assume that in the last two events, $(1-a)\epsilon/2M\pi_j \le 1$ (otherwise

the probabilities of the said events would be zero in view of $|L_n^{R'j} - L_n^j| \le 1$). But

$$\{|\pi_{jn} - \pi_j| \le a\epsilon/2M \ , \pi_j \ge (1-a)\epsilon/2M\} \subseteq \{N_{jn} \ge n\epsilon(1-2a)/2M\}$$

which explains the last step in the chain of inequalities. Next,

$$P\{N_{jn} \ge n\epsilon(1-2a)/2M \ , |L_n^{R'j} - L_n^j| \ge (1-a)\epsilon/2M\pi_j\}$$

$$\le \operatorname*{ess\,sup}_{N_{jn} \ge n\epsilon(1-2a)/2M} P\{|L_n^{R'j} - L_n^j| \ge (1-a)\epsilon/2M\pi_j \mid N_{jn}\}.$$

We will show that for all $\beta > 0$ and $k > 0$, k integer,

$$P\{|L_n^{R'j} - L_n^j| \ge \beta \mid N_{jn} = k\} \le 4\, s(\mathcal{A}^{2M}, 2k)\, e^{-k\beta^2/2^{2M+3}}$$

where $\mathcal{A}^{2M}$ is the class of all $2M$-fold intersections of open or closed linear halfspaces in $\mathbb{R}^{m'}$ and where the function s is defined in section 3.2. In particular, from lemmas 3.2 and 3.6 we know that

$$s(\mathcal{A}^{2M}, 2k) \le \left(s(\mathcal{A}^1, 2k)\right)^{2M}$$

$$\le \left(1 + (2k)^{m'+1}\right)^{2M} \le (1+2k)^{2(m'+1)M}.$$

Combining bounds, with $\beta = (1-a)\epsilon/2M\pi_j$ and $k = (1-2a)n\epsilon/2M$, yields, upon noting that $\pi_j \le 1$ for all j,

$$P\{|L_n^R - L_n| \ge \epsilon\} \le 2e^{-n\epsilon^2/2} + 2Me^{-na^2\epsilon^2/2M^2}$$

$$+ 4M\left(1+(1-2a)n\epsilon/M\right)^{2(m'+1)M} e^{-\left(\frac{(1-2a)n\epsilon}{2M}\right)\left(\frac{(1-a)\epsilon}{2M}\right)^2/2^{2M+3}}$$

from which theorem 6.8 follows if we let $a = 1/3$.

We need only show that for all $\beta > 0$ and $k > 0$, k integer,

$$P\{|L_n^{R'j} - L_n^j| \ge \beta \mid N_{jn} = k\} \le 4\, s(\mathcal{A}^{2M}, 2k) e^{-k\beta^2/2^{2M+3}}.$$

Let

$$w_j^n \varphi(x) = \sum_{i=0}^{m'} w_{ji}^n(V_n)\, \varphi_i(x) \quad , 1 \le j \le M \quad , x \in \mathbb{R}^m ,$$

and let μ_{1k} denote the empirical measure based on the k observations X_i for which $\theta_i = 1$. For given V_n, define the following subsets of $\mathbb{R}^m$,

$$A_0 = \bigcup_{j=2}^{M} \{x : w_1^n \varphi(x) < w_j^n \varphi(x)\}$$

and, with $0 \le N < M$ and $\{i_1, \ldots, i_N\} \subseteq \{2, \ldots, M\}$,

$$A_{\{i_1,\ldots,i_N\}} = \{x : w_1^n \varphi(x) = w_{i_1}^n \varphi(x) = \ldots = w_{i_N}^n \varphi(x) > \sup_{j \notin \{1, i_1, \ldots, i_N\}} w_j^n \varphi(x)\} .$$

Then,

$$L_n^1 - L_n^{R'1} = \mu_1(A_0) - \mu_{1k}(A_0)$$

$$+ \sum_{\substack{\text{all subsets } J \text{ of} \\ \{2,\ldots,M\}}} \left(\int_{A_J} (1 - \xi_{n1}(J^*, x))\, \mu_1(dx) - \int_{A_J} (1 - \xi_{n1}(J^*, x))\, \mu_{1k}(dx) \right)$$

$$= \mu_1(A_0) - \mu_{1k}(A_0)$$

$$- \sum_{\substack{\text{all subsets } J \text{ of} \\ \{2,\ldots,M\}}} (\mu_1(A_J) - \mu_{1k}(A_J))\, \xi_{n1}(J^*, 0)$$

where J^* is the subset obtained by taking the union of $\{1\}$ and $\{i_1, \ldots, i_N\}$, the J-th subset. Notice that we used the fact that $\xi_n(J, x) = \xi_n(J, 0)$ for all x and all J. So,

$$|L_n^1 - L_n^{R'1}| \le |\mu_1(A_0) - \mu_{1k}(A_0)|$$

$$+ 2^{M-1} \sup_{\substack{\text{all subsets } J \text{ of} \\ \{2,\ldots,M\}}} |\mu_1(A_J)-\mu_{1k}(A_J)| .$$

Because $|\mu_1(A_0)-\mu_{1k}(A_0)| = |\mu_1(A_0^c)-\mu_{1k}(A_0^c)|$ and A_0^c is a set from $\mathcal{K}^{M-1} \subseteq \mathcal{K}^{2M}$, and because for all subsets J from $\{2,\ldots,M\}$, with $J=\{i_1,\ldots,i_N\}$,

$$A_J = \bigcap_{j \notin J^*} \{x: w_1^n\varphi(x) > w_j^n\varphi(x)\} \cap \bigcap_{\ell \in J} \{x: w_1^n\varphi(x) \geq w_\ell^n\varphi(x)\}$$

$$\cap \bigcap_{\ell \in J} \{x: w_1^n\varphi(x) \leq w_\ell^n\varphi(x)\}$$

is a set from $\mathcal{K}^{M-N-1+2N} = \mathcal{K}^{M+N-1} \subseteq \mathcal{K}^{2(M-1)} \subseteq \mathcal{K}^{2M}$, we thus have for all V_n that

$$|L_n^1 - L_n^{R'1}| \leq 2^M \sup_{A \in \mathcal{K}^{2M}} |\mu_1(A)-\mu_{1k}(A)|$$

and, by (3.15),

$$P\{|L_n^1 - L_n^{R'1}| \geq \beta\} \leq 4\, s(\mathcal{K}^{2M}, 2k)\, e^{-k(\beta/2^M)^2/8} .$$

The proof of the last inequality of section 6.4 is similar if one replaces ϵ by 2ϵ and omits the term $2e^{-n\epsilon^2/2}$ in the bound. In fact, we obtain the intermediary result

$$P\{|L_n - L_n^{R'}| \geq \epsilon\} \leq 2M\, e^{-2na^2\epsilon^2/M^2}$$

$$+ 4M\left(1+2(1-2a)n\epsilon/M\right)^{2(m'+1)M} e^{-\left(\frac{(1-2a)n\epsilon}{M}\right)\left(\frac{(1-a)}{M}\right)^2/2^{2M+3}} .$$

Q.E.D.

Proof of theorems 6.11 and 6.12

Theorems 6.11 and 6.12 follow immediately from theorems 6.1 and 6.2, from the symmetry of δ_n and $\tilde{\delta}_n$, from

$$E\{|\delta_{n\theta}(V'_n,X)-\delta_{n\theta}(T'_n,X)|\} \le P\{\delta_{n\theta}(V'_n,X)\neq\tilde{\delta}_{n\theta}(T'_n,X)\},$$

and lemma 6.4 given below.

Lemma 6.4. Let M=2 and let $\{\delta_n\}$ be a $\{k_n\}$-nearest neighbor rule. Then, with the holdout decision function $\tilde{\delta}_n$,

$$P\{\delta_{n\theta}(V'_n,X)\neq\tilde{\delta}_{n\theta}(T'_n,X)\} \le (4/\sqrt{\pi})s_n\sqrt{k_n}/n$$

and with the deleted decision function $\tilde{\delta}_n$,

$$P\{\delta_{n\theta}(V'_n,X)\neq\tilde{\delta}_{n\theta}(V'_{n1},X)\} \le (4/\sqrt{\pi})\sqrt{k_n}/n .$$

Proof of lemma 6.4

We first show that if Y is a binomial random variable with parameters n and $\frac{1}{2}$, then for any integer $a\ge 1$,

$$P\{|Y-n/2|\le a/2\} \le \begin{cases}(a+1)2/\sqrt{2\pi n} & , \text{ n even}\\ a\,2\sqrt{2}/\sqrt{2\pi n} & , \text{ n odd}\end{cases}$$
$$\le 2(a+1)/\sqrt{\pi n} \le 4a/\sqrt{\pi n} .$$

Indeed, if n is even, $n\ge 2$, then the central term in the binomial expansion is

$$2^{-n}\binom{n}{n/2} \le (2/\sqrt{2\pi n})\, e^{(1/12n-2/(6n+1))} < 2/\sqrt{2\pi n}$$

by Feller's approximation for n! (Feller, 1968). If n is odd and $n\ge 3$, then the maximal term in the binomial expansion is

$$2^{-n}\binom{n}{\frac{n-1}{2}} < 2^{-(n-1)}\binom{n-1}{\frac{n-1}{2}} < 2/\sqrt{2\pi(n-1)} < 2/\sqrt{\pi n} .$$

This proves the aforementioned inequality.

Now, let M=2, let $(X_1^x,\theta_1^x,Z_1^x),\ldots,(X_n^x,\theta_n^x,Z_n^x)$ be the ordered permutation of V'_n where $x\in\mathbb{R}^m$ and let

$$N_{jn}^{x} = \sum_{i=1}^{k_n} I_{\{\theta_i^x = 1\}}, \quad j=1,2.$$

Then,

$$P\{\delta_{n\theta}(V_n', X) \neq \tilde{\delta}_{n\theta}(T_n', X)\}$$

$$\leq \sup_x \sum_{j=1}^{s_n} P\{\sum_{i=1}^{s_n} \sum_{\ell=1}^{k_n} I_{\{Z_{n-s_n+i} = Z_\ell^x\}} = j, \ |N_{1n}^x - N_{2n}^x| \leq j\}$$

$$\leq \sup_x \sum_{j=1}^{s_n} \frac{\binom{k_n}{j}\binom{n-k_n}{s_n-j}}{\binom{n}{s_n}} \ 4j/\sqrt{\pi k_n}$$

where $\binom{k_n}{j} = 0$ if $j > k_n$. To see this, notice that, conditioned on $(X_{k_n+1}^x, \theta_{k_n+1}^x, Z_{k_n+1}^x)$, the events $\{\sum_{i=1}^{s_n} \sum_{\ell=1}^{k_n} I_{\{Z_{n-s_n+i} = Z_\ell^x\}} = j\}$ and $\{|N_{1n}^x - N_{2n}^x| \leq j\}$ are independent. Also, N_{1n}^x is conditionally binomial with unknown probability parameter p and known counting parameter k_n. It is clear that for all $j \geq 1$,

$$P\{|N_{1n}^x - n/2| \leq j/2 \ | X_{k_n+1}^x, \theta_{k_n+1}^x, Z_{k_n+1}^x\} \leq 4j/\sqrt{\pi k_n}$$

by considering the worst case, that is, $p = \frac{1}{2}$. The probability of the former event is the probability that out of an urn with n balls, k_n black ones and $n-k_n$ red ones, we pick, without replacement, exactly j black balls and $s_n - j$ red balls.

From the properties of the hypergeometric distribution (Roussas, 1973) we know that

$$\sum_{j=1}^{s_n} \binom{k_n}{j}\binom{n-k_n}{s_n-j} \binom{n}{s_n}^{-1} j = k_n s_n / n .$$

Thus,

$$P\{\delta_{n\theta}(V'_n, X) \neq \tilde{\delta}_{n\theta}(T'_n, X)\} \leq (4/\sqrt{\pi})\, s_n \sqrt{k_n}/n .$$

The second part of lemma 6.4 follows at once by letting $s_n = 1$.

Q.E.D.

Proof of theorem 6.13

Let M=2, let $\pi_1 = \pi_2 = \frac{1}{2}$ and let μ_1 and μ_2 both put mass 1 at 0. Let $\{\delta_n\}$ be a $\{k_n\}$-nearest neighbor rule with $k_n = n$ for all n. Let N_{1n} be the number of observations X_i for which $\theta_i = 1$ and let n be even. It is clear that $L_n = \frac{1}{2}$ and that $L_n^D = 1$ if $N_{1n} = \frac{1}{2}$. Thus,

$$E\{(L_n - L_n^D)^2\} \geq (1/2)^2 P\{N_{1n} = n/2\}$$

$$= \frac{1}{4}\binom{n}{n/2} 2^{-n} > 1/4\sqrt{2n}$$

and

$$P\{|L_n - L_n^D| \geq \tfrac{1}{2}\} \geq P\{N_{1n} = n/2\} > 1/\sqrt{2n} .$$

Q.E.D.

Proof of theorem 6.15

Let M=2 and let $\{\delta_n\}$ be a $\{k_n\}$-nearest neighbor rule with a corresponding deleted decision function $\tilde{\delta}_n, n \geq 1$. Let θ_{V_n, X_i}, θ_{V_{ni}, X_i}, $1 \leq i \leq n$, be independent random variables (conditioned on V_n) with

$$P\{\theta_{V_n, X_i} = j \mid V_n\} = \delta_{nj}(V_n, X_i) , \quad 1 \leq j \leq M,$$

and

$$P\{\theta_{V_{ni}, X_i} = j \mid V_n\} = \tilde{\delta}_{nj}(V_{ni}, X_i) , \quad 1 \leq j \leq M.$$

Then,

$$|L_n^R - L_n| \le |L_n^D - L_n| + \left| \left(\sum_{i=1}^{n} (I_{\{\theta_{V_n}, X_i \neq \theta_i\}} - I_{\{\theta_{V_{ni}}, X_i \neq \theta_i\}}) \right) / n \right|$$

and

$$\begin{aligned} & E\{(L_n^R - L_n)^2\} \\ & \le 2E\{(L_n^D - L_n)^2\} + 2E\{(I_{\{\theta_{V_n}, X_1 \neq \theta_1\}} - I_{\{\theta_{V_{n1}}, X_1 \neq \theta_1\}})^2\} \\ & \le 2(1+6k_n)/n + 2P\{\theta_{V_n}, X_1 \neq \theta_{V_{n1}}, X_1\} \\ & \le 2(1+6k_n)/n + 2 \sup_x P\{|N_{1n}^x - N_{2n}^x| \le 1\} \\ & \le 2(1+6k_n)/n + 8/\sqrt{\pi k_n} \end{aligned}$$

where we used (6.43) and the bound derived in the proof of lemma 6.4, and where

$$N_{jn}^x = \sum_{i=1}^{k_n} I_{\{\theta_i^x = j\}}, j=1,2.$$

If theorem 6.11 is used instead of (6.43), then the factor $6k_n$ in the bound can be replaced by $24\sqrt{k_n/\pi}$.

Q.E.D.

BIBLIOGRAPHY

G.BENNETT (1962).Probability inequalities for the sum of independent random variables.Journal of the American Statistical Association 57, 33-45.

L.BREIMAN(1968).Probability.Addison-Wesley,Reading,Mass.

T.CACOULLOS(1965).Estimation of a multivariate density.Annals of the Institute of Statistical Mathematics 18,179-190.

K.L.CHUNG(1968).A Course in Probability Theory. Harcourt,Brace and World,Inc.,N.Y..

T.M.COVER(1965).Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition.IEEE Transactions on Electronic Computers 10,326-334.

T.M.COVER(1968).Rates of convergence of nearest neighbor decision procedures.Proceedings of the First Annual Hawaii Conference on Systems Theory,413-415.

T.M.COVER(1969).Learning in pattern recognition.Methodologies of Pattern Recognition.S.Watanabe Ed.,Academic Press,N.Y.,111-132.

T.M.COVER,P.E.HART(1967).Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13,21-27.

T.M.COVER,T.J.WAGNER(1975).Topics in statistical pattern recognition.Communication and Cybernetics 10,1-16.

L.P.DEVROYE,T.J.WAGNER(1976).A distribution free performance bound in error estimation.To appear in IEEE Transactions on Information Theory 22.

S.W.DHARMADHIKARI,K.JOGDEO(1969).Bounds on moments of certain random variables.Annals of Mathematical Statistics 40,1506-1509.

A.DVORETZKY,J.KIEFER,J.WOLFOWITZ(1956).Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator.Annals of Mathematical Statistics 27,642-669.

R.O.DUDA,P.E.HART(1973).Pattern Classification and Scene Analysis. Wiley,N.Y..

W.FELLER(1966).An Introduction to Probability Theory and its Applications vol.2.John Wiley,N.Y..

W.FELLER(1968).An Introduction to Probability Theory and its Applications vol.1.John Wiley,N.Y..

T.FERGUSON(1967).Mathematical Statistics.A Decision Theoretic Approach.Academic Press,N.Y..

E.FIX,J.L.HODGES(1951).Discriminatory Analysis.Nonparametric Discrimination:Consistency Properties. Report 4,Project No.21-49-004, USAF School of Aviation Medicine,Randolph Field,Texas.

A.FOLDES,P.REVESZ(1974).A general method for density estimation. Studia Scientiarium Mathematicarum Hungarica 9,81-92.

D.K.FUK,S.V.NAGAEV(1971).Probability inequalities for sums of independent random variables.Theory of Probability and its Applications 16,643-660.

K.FUKUNAGA(1972).Introduction to Statistical Pattern Recognition. Academic Press,N.Y..

A.M.GARSIA(1970).Topics in Almost Everywhere Convergence.Markham Publishing Co.,Chicago.

N.GLICK(1974).Consistency conditions for probability estimators and integrals of density estimators.Utilitas Mathematica 6,61-74.

P.E.HART(1968).The condensed nearest neighbor rule.IEEE Transactions on Information Theory 14,515-516.

E.HERTZ(1969).On the convergence rates in the central limit theorem. Annals of Mathematical Statistics 40,475-479.

Y.C.HO,A.K.AGRAWALA(1968).On pattern classification algorithms: introduction and survey.IEEE Transactions on Automatic Control 13, 676-690.

W.HOEFFDING(1963).Probability inequalities for sums of bounded random variables.Journal of the American Statistical Association 58, 13-30.

J.KIEFER(1961).On large deviations of the empiric d.f. of vector chance variables.Annals of Mathematical Statistics 32,649-660.

J.KIEFER,J.WOLFOWITZ(1958).On the deviations of the empiric distribution function of vector chance variables.Transactions of the American Mathematical Society 87,173-186.

B.K.KIM,J.VAN RYZIN(1975).Uniform consistency of a histogram density estimator and modal estimation.Communications in Statistics 4,303-315.

P.A.LACHENBRUCH(1967).An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis.Biometrics 23,639-645.

M.LOEVE (1963).Probability Theory. Van Nostrand,Princeton,N.J..

D.O.LOFTSGAARDEN,C.P.QUESENBERRY(1965).A nonparametric estimate of a multivariate density function.Annals of Mathematical Statistics 36,1049-1051.

D.S.MITRINOVIC(1964).Elementary Inequalities. P.Noordhoff Ltd., Groningen,Netherlands.

D.S.MOORE,E.G.HENRICHON(1969).Uniform consistency of some estimates of a density function.Annals of Mathematical Statistics 40, 1499-1502.

D.S.MOORE,J.W.YACKEL(1975).Consistency Properties of Nearest Neighbor Density Function Estimators.Dept.of Statistics,Division of Math.Sciences,Mimeograph Series 426,Purdue Univ.,Lafayette,Ind..

E.A.NADARAYA(1964).On estimating regression.Theory of Probability and its Applications 9,141-142.

E.A.NADARAYA(1965).On nonparametric estimates of density functions

and regression curves. Theory of Probability and its Applications 10, 186-190.

E.A.NADARAYA(1970). Remarks on nonparametric estimates for density functions and regression curves. Theory of Probability and its Applications 15, 134-137.

E.A.NADARAYA(1973). On the convergence in the L_2 norm of probability density estimates. Theory of Probability and its Applications 18, 808-811.

G.NAGY(1968). State of art in pattern recognition. Proceedings of the IEEE 56, 836-862.

L.V.OSIPOV, V.V.PETROV(1967). On an estimate of the remainder term in the central limit theorem. Theory of Probability and its Applications 12, 281-286.

E.PARZEN(1962). On the estimation of a probability density function and the mode. Annals of Mathematical Statistics 33, 1065-1076.

E.A.PATRICK, F.P.FISCHER III(1970). A generalized k-nearest neighbor rule. Information and Control 16, 128-152.

J.PRATT(1959). On a general concept of "in probability". Annals of Mathematical Statistics 30, 549-558.

J.PRATT(1960). On interchanging limits and integrals. Annals of Mathematical Statistics 31, 74-77.

R.RANGA RAO(1962). Relations between weak and uniform convergence of measures with applications. Annals of Mathematical Statistics 33, 659-680.

W.H.ROGERS, T.J.WAGNER(1976). A finite sample distribution free performance bound for local discrimination rules. To appear in The Annals of Statistics.

B.ROSEN(1970). On bounds on the central moments of even order of a sum of independent random variables. Annals of Mathematical Sta-

tistics 41,1074-1077.

M.ROSENBLATT(1957).Remarks on some nonparametric estimates of a density function.Annals of Mathematical Statistics 27,832-837.

G.ROUSSAS(1973).A First Course in Mathematical Statistics. Addison-Wesley,Reading,Mass..

H.L.ROYDEN(1968).Real Analysis. Macmillan,London.

E.F.SCHUSTER(1969).Estimation of a probability density function and its derivatives.Annals of Mathematical Statistics 40,1187-1195.

G.SEBESTYEN(1962).Decision Making Processes in Pattern Recognition.Macmillan,N.Y..

R.J.SERFLING(1974).Probability inequalities for the sum in sampling without replacement.Annals of Statistics 2,39-48.

C.J.STONE(1976).Nonparametric Regression and its Applications. Manuscript,Univ.of California at Los Angeles.

G.TOUSSAINT(1974).Bibliography on estimation of misclassification. IEEE Transactions on Information Theory 20,472-479.

J.VAN RYZIN(1966).Bayes risk consistency of classification procedures using density estimation.Sankhya Ser.A 28,261-270.

J.VAN RYZIN(1969).On strong consistency of density estimates. Annals of Mathematical Statistics 40,1765-1772.

V.N.VAPNIK,A.Ya.CHERVONENKIS(1971).On the uniform convergence of the relative frequencies of events to their probabilities.Theory of Probability and its Applications 16,264-280.

T.J.WAGNER(1971).Convergence of the nearest neighbor rule.IEEE Transactions on Information Theory 17,566-571.

T.J.WAGNER(1973a).Deleted estimates of the Bayes risk.Annals of Statistics 1,359-362.

T.J.WAGNER(1973b).Strong consistency of a nonparametric estimate of a density function.IEEE Transactions on Systems,Man and Cyber-

netics 3,289-290.

G.WAHBA(1973).Interpolating spline methods for density estimation. II.Variable knots.Technical Rept.337,Dept.of Statistics,Univ.of Wisconsin,Madison,Wisconsin.

G.WAHBA(1975a).Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation.Annals of Statistics 3,15-29.

G.WAHBA(1975b).Interpolating spline methods for density estimation. I.Equi-spaced knots.Annals of Statistics 3,30-44.

E.J.WEGMAN(1972).Nonparametric probability density estimation.I. A summary of available methods.Technometrics 14,533-546.

P.WHITTLE(1960).Bounds for the moments of linear and quadratic forms in independent variables.Theory of Probability and its Applications 5,302-305.

D.L.WILSON(1972).Asymptotic properties of nearest neighbor rules using edited data.IEEE Transactions on Systems,Man and Cybernetics 2,408-421.

J.WOLFOWITZ(1960).Convergence of the empiric distribution function on half-spaces.Contributions in Probability and Statistics.Edited by I.Olkin et al.,Stanford Univ.Press,Stanford ,Cal.,504-507.

V.M.ZOLOTAREV(1966).An absolute estimate of the remainder term in the central limit theorem.Theory of Probability and its Applications 11,95-105.

VITA

Luc P J A Devroye was born in Tienen, Belgium, on August, 6, 1948. He obtained the Master of Science degree from the Department of Electrical Engineering, Catholic University of Louvain, Belgium, in 1971. From 1971 to 1972 he was with the Electronics Research Laboratory at the same University supported by a research grant from the IWONL. From 1972 to 1974 he was, supported by a Japanese Government Scholarship, working with Professor K. Fujii at the University of Osaka, Japan, where he was engaged in the analysis of probabilistic automata and probabilistic search techniques. In 1974 he joined the University of Texas at Austin where he has completed his work for the Ph.D. degree in Electrical Engineering under the supervision of Professor T.J. Wagner. His current research interests are in the fields of probabilistic optimization methods and statistical pattern recognition.

Permanent address : 1405 W. North Loop # 209
Austin, Texas 78756

This dissertation was typed by Nancy Boster.